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

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

Last edited November 27, 2023

Contents

1 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:

Variable Names	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	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	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

1.1 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.

1.2 Write Up:

1.2.1 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"?

1.2.2 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

1.2.3 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.

1.2.4 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.

2 Import Data

```
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_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
```

```
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

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

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

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

2.0.0.0.1 DATA CLEANING 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))
```

```
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.

```
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")))
```

```
#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 = ""))
}
```

2.0.0.0.2 MISSING DATA ANS IMPUTATION

```
## [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
```

```
#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.

```
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
```

###Exploratory Data analysis

Summary statistics for the numeric variables:

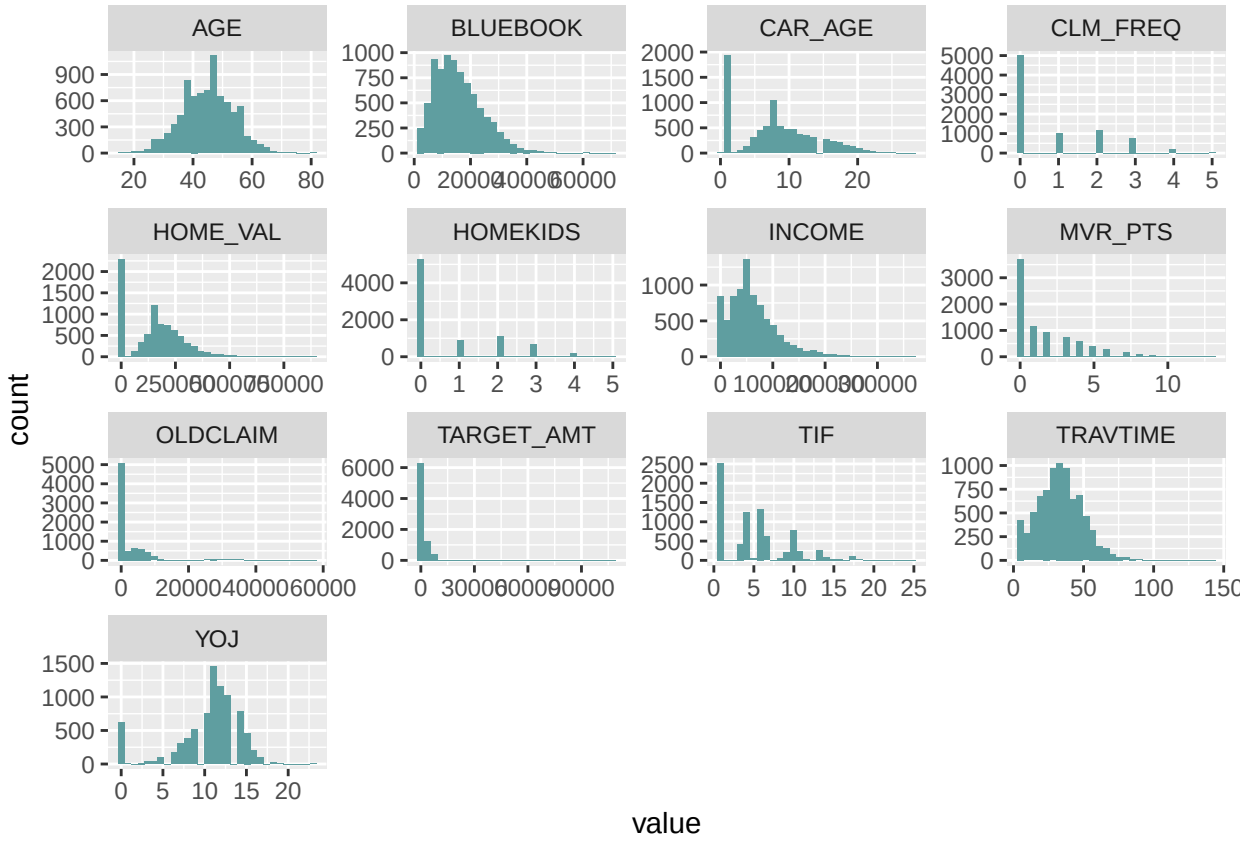
```
df_insur_train %>%
  select(TARGET_AMT, AGE, YOJ, INCOME, HOME_VAL, TRAVTIME, BLUEBOOK, TIF,
         OLDCLAIM, CLM_FREQ, MVR_PTS, CAR_AGE, HOMEKIDS) %>%
  describe()
```

```
##            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.

```
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`.
```

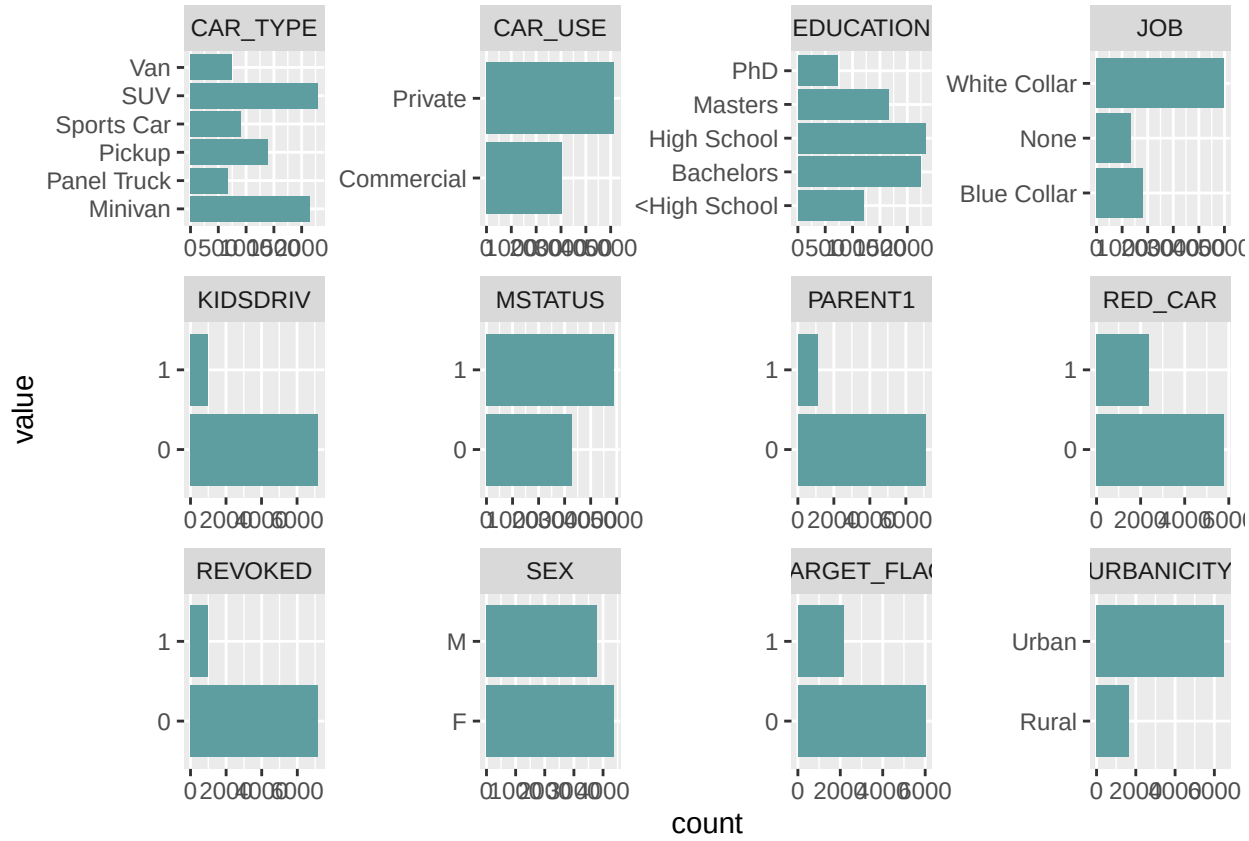


The histogram for TARGET_AMT, CAR_AGE, CLM_FREQ,HOME_VAL, INCOME, MRV_PTS, OLDCLAIM, and TIF are clearly not normally distributed and will need to be transformed if the residuals are not normally distributed.

We will explore the proportions of the discrete variables.

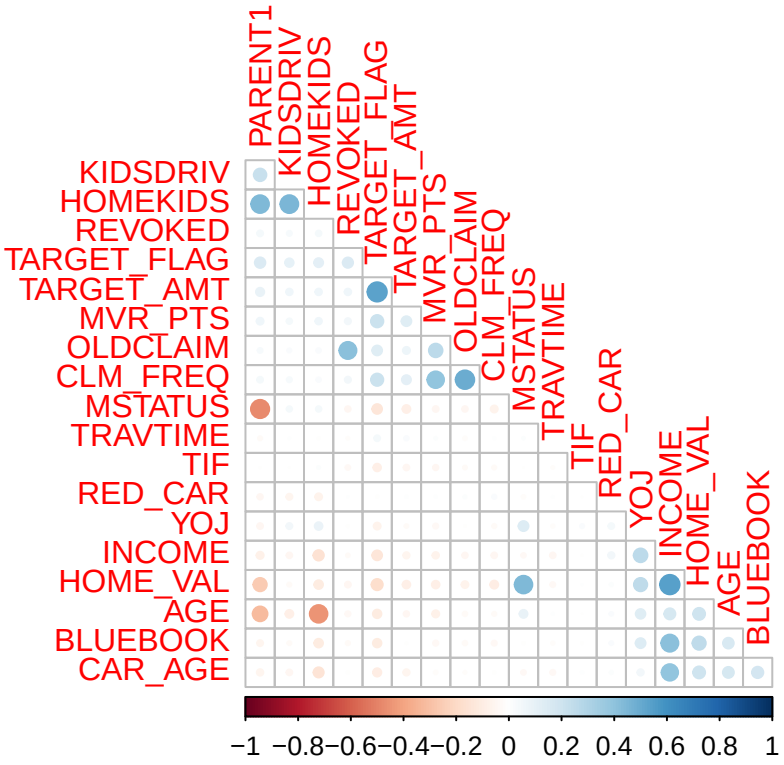
```
df_insur_train %>%
  select(TARGET_FLAG, KIDSDRIV, PARENT1, MSTATUS, SEX, EDUCATION,
         JOB, CAR_USE, CAR_TYPE, RED_CAR, REVOKED, URBANICITY) %>%
  gather() %>%
  ggplot(aes(x = value)) +
  geom_bar(fill = "cadetblue") +
  coord_flip()+
  facet_wrap(~key, scales = "free")
```

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



To check for coliniearity through the corellation 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.

2.0.1 Logistic Regression Model

```
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 degree of freedom adjusted variance inflation factors suggests that there is no concerning colinearity becasue all of the values are less than 3.

```
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 < 0.0000000000000022 ***
## - CAR_TYPE    5   7441.4 7495.4  88.91 < 0.0000000000000022 ***
## - CAR_USE     1   7437.2 7499.2  84.65 < 0.0000000000000022 ***
## - REVOKED     1   7450.3 7512.3  97.79 < 0.0000000000000022 ***
## - URBANICITY  1   7967.7 8029.7 615.20 < 0.0000000000000022 ***
## ---
## 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 < 0.0000000000000022 ***
## - CAR_TYPE    5   7441.6 7493.6  89.00 < 0.0000000000000022 ***
## - CAR_USE     1   7437.2 7497.2  84.65 < 0.0000000000000022 ***
```

```
## - 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
```

```
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  0.00000000006611971
## 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
```

The variable 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 variable 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

2.0.2 ASSESING MODEL PERFORMANCE

We are going to first predict the probabilities of a car crash using the final backward stepwise regression model from which we will then call the predicted car crash based on the probability of 0.5.

```
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.

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

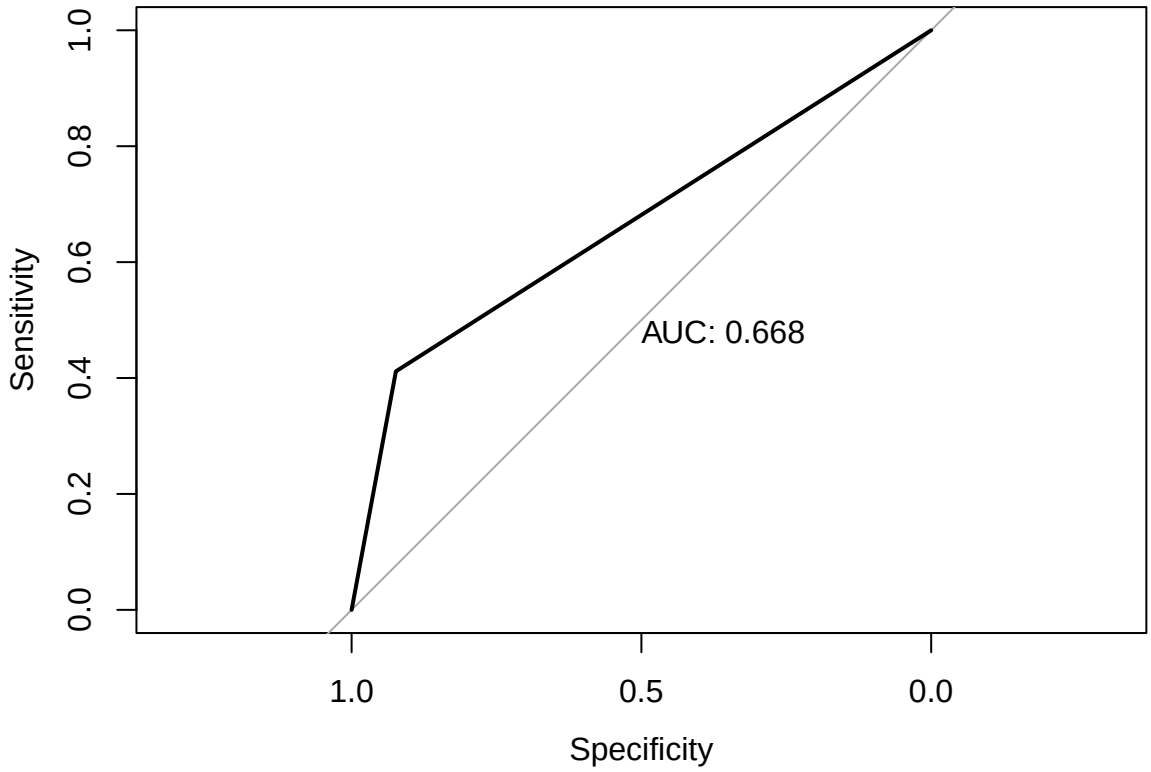
```
## Setting levels: control = 0, case = 1
```

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

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

```
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.au
##
## 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")
```

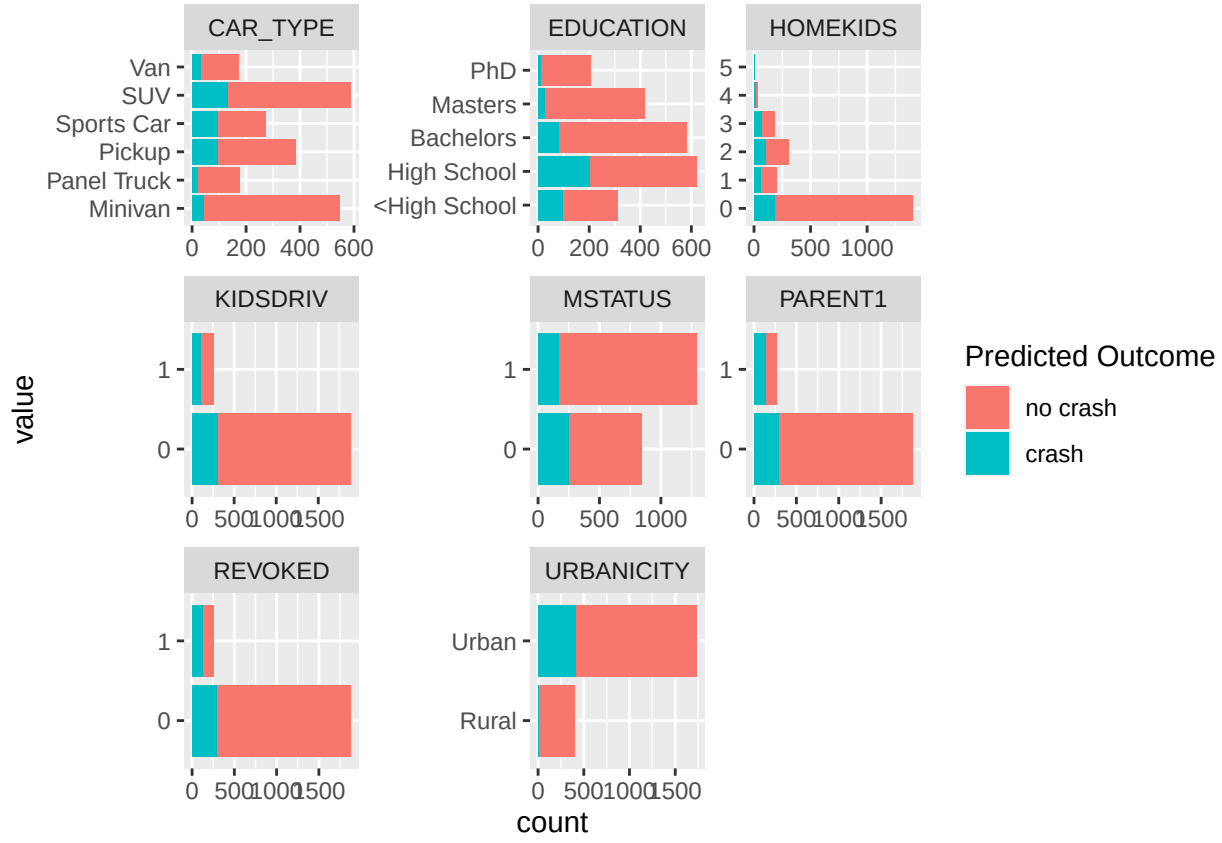
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5549 1267
##          1  459  886
##
##                Accuracy : 0.7885
##                  95% CI : (0.7795, 0.7973)
##     No Information Rate : 0.7362
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   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.

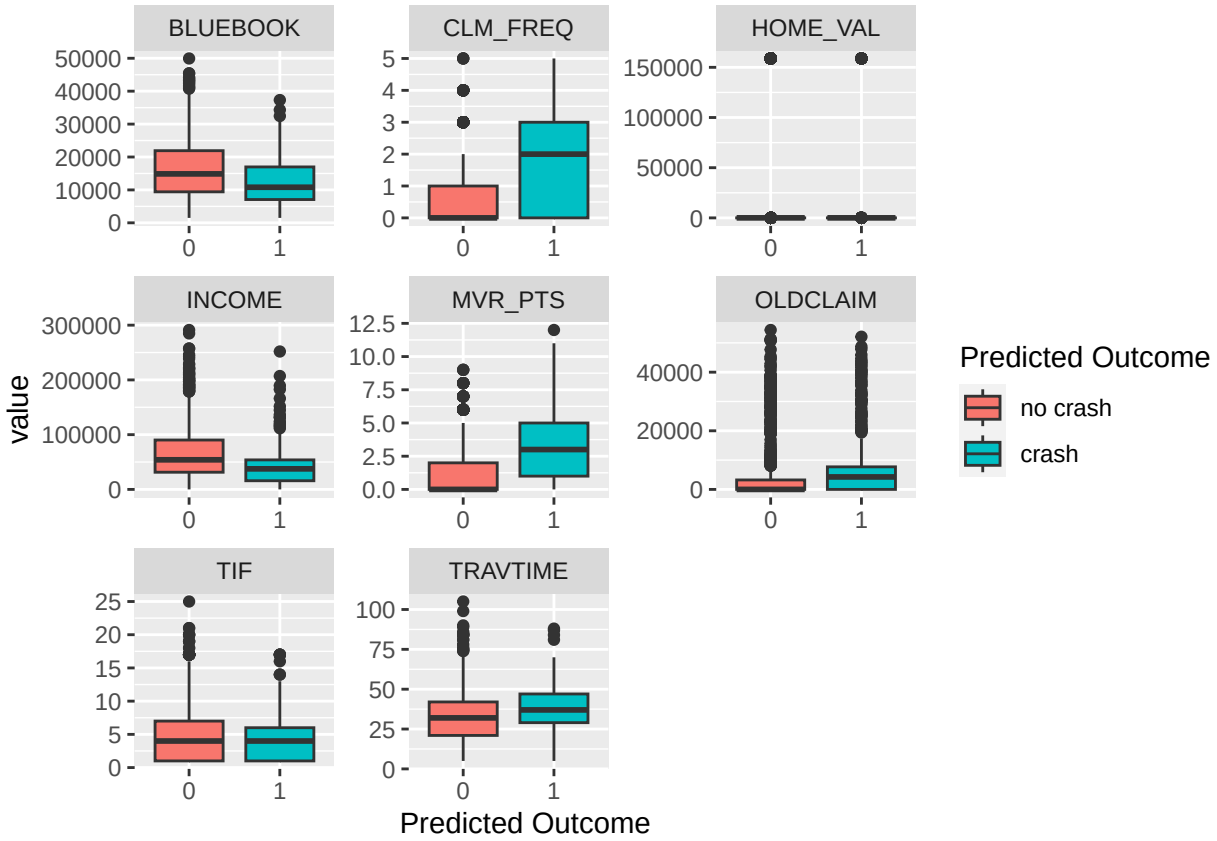
```
df_insur_eval$log_pred_prob <- predict(log_step,
                                       newdata = df_insur_eval[,-c(1:2)],
                                       type = "response")
df_insur_eval$log_pred <- ifelse(df_insur_eval$log_pred_prob > 0.5, 1, 0)
```

```
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")
```

2.0.2.1 PREDICTING CAR CRASHES WITH THE EVALUATIONS DATASET




```
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 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.

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

2.0.2.2 MULTIPLE LINEAR REGRESSION

```
##
## 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
```

```
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
```

- The degree of freedom adjusted variance inflation factors suggests that there is no concerning colinearity because all of the values are less than 3.

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

```
## Start:  AIC=137507.2
## 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)
## - EDUCATION   4  112973325 168341925971 137505   1.3647              0.243461
## - JOB         2   37175115 168266127761 137505   0.8982              0.407355
## - YOJ         1     721983 168229674629 137505   0.0349              0.851837
## - RED_CAR     1    2943744 168231896390 137505   0.1422              0.706069
## - AGE         1    4070588 168233023234 137505   0.1967              0.657414
## - HOME_VAL    1   12996862 168241949508 137506   0.6280              0.428105
## - HOMEKIDS    1   24557178 168253509825 137506   1.1866              0.276043
## <none>                     168228952646 137507
## - OLDCLAIM    1   46092130 168275044777 137507   2.2272              0.135636
## - BLUEBOOK    1   54188751 168283141398 137508   2.6185              0.105666
## - SEX         1   73160540 168302113186 137509   3.5352              0.060115
## - CAR_AGE     1   91935551 168320888198 137510   4.4424              0.035087
## - PARENT1     1  140176049 168369128695 137512   6.7735              0.009269
## - CLM_FREQ    1  144985337 168373937983 137512   7.0058              0.008140
## - INCOME      1  199498442 168428451088 137515   9.6400              0.001910
## - REVOKED     1  226653426 168455606072 137516  10.9521              0.000939
## - KIDSDRIV    1  250471148 168479423794 137517  12.1030              0.000506
## - TRAVTIME    1  315547889 168544500535 137521  15.2476     0.000095059126720
## - TIF         1  320673202 168549625848 137521  15.4953     0.000083402066389
## - MSTATUS     1  348294698 168577247344 137522  16.8300     0.000041281622931
## - CAR_TYPE    5  588299169 168817251815 137526   5.6854     0.000030632987734
## - CAR_USE     1  540457973 168769410619 137531  26.1155     0.000000328913631
## - MVR_PTS     1 1032912847 169261865493 137555  49.9114     0.000000000001741
```

```
## - URBANICITY  1 2629760500 170858713146 137632 127.0728 < 0.00000000000000022
##
## - EDUCATION
## - JOB
## - YOJ
## - RED_CAR
## - AGE
## - HOME_VAL
## - HOMEKIDS
## <none>
## - OLDCLAIM
## - BLUEBOOK
## - SEX        .
## - CAR_AGE    *
## - PARENT1    **
## - CLM_FREQ   **
## - INCOME     **
## - REVOKED    ***
## - KIDSDRIV   ***
## - TRAVTIME   ***
## - TIF        ***
## - MSTATUS    ***
## - CAR_TYPE   ***
## - CAR_USE    ***
## - MVR_PTS    ***
## - URBANICITY ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=137504.7
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + 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)
## - YOJ         1     715499 168342641471 137503   0.0346            0.8525095
## - RED_CAR     1    3692846 168345618818 137503   0.1784            0.6727542
## - AGE         1    4530249 168346456220 137503   0.2189            0.6399170
## - JOB         2   53182904 168395108875 137503   1.2847            0.2767907
## - HOME_VAL    1   19095874 168361021845 137504   0.9226            0.3368304
## - HOMEKIDS    1   28954388 168370880359 137504   1.3989            0.2369499
## <none>                     168341925971 137505
## - OLDCLAIM    1   44499076 168386425047 137505   2.1499            0.1426219
## - BLUEBOOK    1   49667832 168391593804 137505   2.3996            0.1214074
## - SEX         1   73285647 168415211618 137506   3.5406            0.0599193
## - PARENT1     1  133424136 168475350107 137509   6.4460            0.0111385
## - CLM_FREQ    1  143955616 168485881587 137510   6.9548            0.0083752
## - CAR_AGE     1  196065758 168537991730 137512   9.4724            0.0020928
## - REVOKED     1  227444003 168569369974 137514  10.9884            0.0009209
## - KIDSDRIV    1  245327046 168587253017 137515  11.8523            0.0005788
## - INCOME      1  251607514 168593533485 137515  12.1558            0.0004919
## - TRAVTIME    1  309045658 168650971629 137518  14.9307            0.0001124
## - TIF         1  316242112 168658168083 137518  15.2784     0.000093523857180
```

```
## - MSTATUS     1  336072100 168677998072 137519  16.2364     0.000056417536029
## - CAR_TYPE    5  603180493 168945106464 137524   5.8282     0.000022219779728
## - CAR_USE     1  502406642 168844332613 137527  24.2725     0.000000852762858
## - MVR_PTS     1 1035925877 169377851848 137553  50.0481     0.000000000001625
## - URBANICITY  1 2632166732 170974092703 137629 127.1663 < 0.00000000000000022
##
## - YOJ
## - RED_CAR
## - AGE
## - JOB
## - HOME_VAL
## - HOMEKIDS
## <none>
## - OLDCLAIM
## - BLUEBOOK
## - SEX        .
## - PARENT1    *
## - CLM_FREQ   **
## - CAR_AGE    **
## - REVOKED    ***
## - KIDSDRIV   ***
## - INCOME     ***
## - TRAVTIME   ***
## - TIF        ***
## - MSTATUS    ***
## - CAR_TYPE   ***
## - CAR_USE    ***
## - MVR_PTS    ***
## - URBANICITY ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=137502.8
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + 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    3720420 168346361891 137501   0.1798            0.6715876
## - AGE         1    4063966 168346705436 137501   0.1964            0.6576843
## - JOB         2   52477469 168395118940 137501   1.2678            0.2815040
## - HOME_VAL    1   19039745 168361681215 137502   0.9200            0.3375128
## - HOMEKIDS    1   28285094 168370926565 137502   1.3667            0.2424169
## <none>                     168342641471 137503
## - OLDCLAIM    1   44784224 168387425694 137503   2.1639            0.1413253
## - BLUEBOOK    1   49507178 168392148648 137503   2.3921            0.1219891
## - SEX         1   73311878 168415953348 137504   3.5423            0.0598584
## - PARENT1     1  133720842 168476362312 137507   6.4611            0.0110443
## - CLM_FREQ    1  144257210 168486898680 137508   6.9702            0.0083035
## - CAR_AGE     1  195557181 168538198651 137510   9.4490            0.0021197
## - REVOKED     1  227788923 168570430393 137512  11.0063            0.0009120
## - KIDSDRIV    1  246477814 168589119285 137513  11.9093            0.0005614
## - INCOME      1  254077597 168596719067 137513  12.2766            0.0004612
```

```
## - TRAVTIME    1  308790877 168651432347 137516  14.9202            0.0001130
## - TIF         1  316479682 168659121153 137516  15.2917     0.000092869169209
## - MSTATUS     1  342498810 168685140280 137517  16.5489     0.000047859362353
## - CAR_TYPE    5  604480764 168947122235 137522   5.8415     0.000021566189006
## - CAR_USE     1  503053229 168845694699 137525  24.3066     0.000000837831139
## - MVR_PTS     1 1037886792 169380528262 137551  50.1487     0.000000000001544
## - URBANICITY  1 2631634650 170974276120 137627 127.1556 < 0.00000000000000022
##
## - RED_CAR
## - AGE
## - JOB
## - HOME_VAL
## - HOMEKIDS
## <none>
## - OLDCLAIM
## - BLUEBOOK
## - SEX        .
## - PARENT1    *
## - CLM_FREQ   **
## - CAR_AGE    **
## - REVOKED    ***
## - KIDSDRIV   ***
## - INCOME     ***
## - TRAVTIME   ***
## - TIF        ***
## - MSTATUS    ***
## - CAR_TYPE   ***
## - CAR_USE    ***
## - MVR_PTS    ***
## - URBANICITY ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=137500.9
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + 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    4205824 168350567714 137499   0.2032            0.6521319
## - JOB         2   52742032 168399103923 137499   1.2743            0.2796751
## - HOME_VAL    1   18639050 168365000941 137500   0.9007            0.3426237
## - HOMEKIDS    1   28094611 168374456501 137500   1.3576            0.2439853
## <none>                     168346361891 137501
## - OLDCLAIM    1   44944146 168391306036 137501   2.1718            0.1405971
## - BLUEBOOK    1   50247645 168396609536 137501   2.4281            0.1192149
## - SEX         1   75531397 168421893288 137503   3.6499            0.0561076
## - PARENT1     1  134030671 168480392562 137505   6.4768            0.0109477
## - CLM_FREQ    1  143794268 168490156159 137506   6.9486            0.0084046
## - CAR_AGE     1  196233754 168542595644 137508   9.4826            0.0020812
## - REVOKED     1  227881680 168574243571 137510  11.0119            0.0009093
## - KIDSDRIV    1  248032927 168594394818 137511  11.9857            0.0005389
## - INCOME      1  254440987 168600802878 137511  12.2953            0.0004565
```

```
## - TRAVTIME    1  308045823 168654407713 137514  14.8857            0.0001151
## - TIF         1  316063783 168662425673 137514  15.2731     0.000093784126533
## - MSTATUS     1  342195425 168688557316 137515  16.5359     0.000048187653859
## - CAR_TYPE    5  606733702 168953095593 137520   5.8638     0.000020506692053
## - CAR_USE     1  502491763 168848853653 137523  24.2819     0.000000848600989
## - MVR_PTS     1 1037805615 169384167506 137549  50.1499     0.000000000001543
## - URBANICITY  1 2629404064 170975765955 137625 127.0607 < 0.00000000000000022
##
## - AGE
## - JOB
## - HOME_VAL
## - HOMEKIDS
## <none>
## - OLDCLAIM
## - BLUEBOOK
## - SEX        .
## - PARENT1    *
## - CLM_FREQ   **
## - CAR_AGE    **
## - REVOKED    ***
## - KIDSDRIV   ***
## - INCOME     ***
## - TRAVTIME   ***
## - TIF        ***
## - MSTATUS    ***
## - CAR_TYPE   ***
## - CAR_USE    ***
## - MVR_PTS    ***
## - URBANICITY ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=137499.1
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + 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)
## - JOB         2   52060918 168402628632 137498   1.2580            0.2842792
## - HOME_VAL    1   17362137 168367929851 137498   0.8391            0.3596893
## - HOMEKIDS    1   23918738 168374486452 137498   1.1559            0.2823414
## <none>                     168350567714 137499
## - OLDCLAIM    1   45005585 168395573299 137499   2.1750            0.1403066
## - BLUEBOOK    1   55758683 168406326397 137500   2.6947            0.1007217
## - SEX         1   80510631 168431078346 137501   3.8909            0.0485823
## - PARENT1     1  130631041 168481198755 137503   6.3131            0.0120041
## - CLM_FREQ    1  144592075 168495159789 137504   6.9878            0.0082224
## - CAR_AGE     1  193136365 168543704079 137506   9.3338            0.0022569
## - REVOKED     1  227253215 168577820929 137508  10.9826            0.0009237
## - INCOME      1  254497608 168605065323 137509  12.2993            0.0004556
## - KIDSDRIV    1  263276075 168613843789 137510  12.7235            0.0003632
## - TRAVTIME    1  308608544 168659176258 137512  14.9143            0.0001134
## - TIF         1  315749371 168666317085 137512  15.2594     0.000094465403665
```

```
## - MSTATUS     1  341996467 168692564181 137514  16.5279     0.000048390802922
## - CAR_TYPE    5  619278163 168969845877 137519   5.9857     0.000015580507719
## - CAR_USE     1  501228464 168851796178 137521  24.2232     0.000000874768177
## - MVR_PTS     1 1034330648 169384898362 137547  49.9868     0.000000000001676
## - URBANICITY  1 2629742633 170980310348 137624 127.0895 < 0.00000000000000022
##
## - JOB
## - HOME_VAL
## - HOMEKIDS
## <none>
## - OLDCLAIM
## - BLUEBOOK
## - SEX        *
## - PARENT1    *
## - CLM_FREQ   **
## - CAR_AGE    **
## - REVOKED    ***
## - INCOME     ***
## - KIDSDRIV   ***
## - TRAVTIME   ***
## - TIF        ***
## - MSTATUS    ***
## - CAR_TYPE   ***
## - CAR_USE    ***
## - MVR_PTS    ***
## - URBANICITY ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=137497.7
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + 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)
## - HOME_VAL    1   14645525 168417274157 137496   0.7077            0.4002198
## - HOMEKIDS    1   24498492 168427127125 137497   1.1839            0.2765995
## <none>                     168402628632 137498
## - OLDCLAIM    1   45411726 168448040358 137498   2.1945            0.1385423
## - BLUEBOOK    1   56472471 168459101103 137498   2.7290            0.0985787
## - SEX         1   84647043 168487275675 137500   4.0905            0.0431565
## - PARENT1     1  128572546 168531201179 137502   6.2132            0.0126997
## - CLM_FREQ    1  144482815 168547111447 137503   6.9821            0.0082487
## - REVOKED     1  226841046 168629469679 137507  10.9620            0.0009340
## - CAR_AGE     1  252930354 168655558987 137508  12.2228            0.0004746
## - KIDSDRIV    1  269041590 168671670222 137509  13.0013            0.0003131
## - INCOME      1  275761351 168678389983 137509  13.3261            0.0002634
## - TRAVTIME    1  315972468 168718601100 137511  15.2693     0.000093976594773
## - TIF         1  318407358 168721035991 137511  15.3869     0.000088312365853
## - MSTATUS     1  350267899 168752896531 137513  16.9266     0.000039236697154
## - CAR_TYPE    5  593304123 168995932755 137516   5.7342     0.000027450912813
## - CAR_USE     1  945364644 169347993276 137541  45.6844     0.000000000014850
## - MVR_PTS     1 1033236320 169435864953 137546  49.9308     0.000000000001724
## - URBANICITY  1 2634462409 171037091042 137622 127.3095 < 0.00000000000000022
```

```
##
## - HOME_VAL
## - HOMEKIDS
## <none>
## - OLDCLAIM
## - BLUEBOOK   .
## - SEX        *
## - PARENT1    *
## - CLM_FREQ   **
## - REVOKED    ***
## - CAR_AGE    ***
## - KIDSDRIV   ***
## - INCOME     ***
## - TRAVTIME   ***
## - TIF        ***
## - MSTATUS    ***
## - CAR_TYPE   ***
## - CAR_USE    ***
## - MVR_PTS    ***
## - URBANICITY ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=137496.4
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + MSTATUS +
##     SEX + 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)
## - HOMEKIDS    1   27287656 168444561813 137496   1.3187             0.2508567
## <none>                    168417274157 137496
## - OLDCLAIM    1   45591268 168462865426 137497   2.2033             0.1377578
## - BLUEBOOK    1   54872492 168472146649 137497   2.6518             0.1034721
## - SEX         1   83429381 168500703538 137498   4.0318             0.0446822
## - PARENT1     1  124419205 168541693362 137500   6.0127             0.0142239
## - CLM_FREQ    1  147827259 168565101416 137502   7.1440             0.0075367
## - REVOKED     1  229392183 168646666340 137505  11.0857             0.0008738
## - CAR_AGE     1  252238978 168669513135 137507  12.1898             0.0004831
## - KIDSDRIV    1  268793490 168686067647 137507  12.9898             0.0003151
## - TRAVTIME    1  318263356 168735537513 137510  15.3805     0.000088611642649
## - TIF         1  318732237 168736006394 137510  15.4032     0.000087557378982
## - CAR_TYPE    5  590986255 169008260413 137515   5.7120     0.000028855156405
## - INCOME      1  492211459 168909485617 137518  23.7868     0.000001096580082
## - MSTATUS     1  566327579 168983601736 137522  27.3686     0.000000172331356
## - CAR_USE     1  949039474 169366313631 137540  45.8637     0.000000000013558
## - MVR_PTS     1 1038292745 169455566902 137545  50.1769     0.000000000001522
## - URBANICITY  1 2628337227 171045611384 137621 127.0181 < 0.00000000000000022
##
## - HOMEKIDS
## <none>
## - OLDCLAIM
## - BLUEBOOK
## - SEX        *
## - PARENT1    *
```

```
## - CLM_FREQ   **
## - REVOKED    ***
## - CAR_AGE    ***
## - KIDSDRIV   ***
## - TRAVTIME   ***
## - TIF        ***
## - CAR_TYPE   ***
## - INCOME     ***
## - MSTATUS    ***
## - CAR_USE    ***
## - MVR_PTS    ***
## - URBANICITY ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=137495.7
## TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS + SEX + 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)
## <none>                     168444561813 137496
## - OLDCLAIM    1   45419325 168489981138 137496   2.1949             0.1385099
## - BLUEBOOK    1   50988789 168495550603 137496   2.4640             0.1165201
## - SEX         1   78268143 168522829956 137497   3.7823             0.0518328
## - CLM_FREQ    1  148061580 168592623393 137501   7.1550             0.0074905
## - REVOKED     1  232739006 168677300819 137505  11.2470             0.0008012
## - PARENT1     1  234989770 168679551584 137505  11.3558             0.0007556
## - CAR_AGE     1  266478380 168711040194 137507  12.8774             0.0003345
## - TIF         1  315163525 168759725338 137509  15.2301     0.000095940549140
## - TRAVTIME    1  316156093 168760717906 137509  15.2781     0.000093539231378
## - KIDSDRIV    1  391998577 168836560390 137513  18.9431     0.000013631838555
## - CAR_TYPE    5  589275928 169033837741 137514   5.6953     0.000029962033942
## - INCOME      1  504788207 168949350020 137518  24.3936     0.000000800908179
## - MSTATUS     1  541122596 168985684409 137520  26.1495     0.000000323183925
## - CAR_USE     1  954849446 169399411259 137540  46.1426     0.000000000011768
## - MVR_PTS     1 1046851930 169491413743 137544  50.5886     0.000000000001236
## - URBANICITY  1 2616514683 171061076497 137619 126.4418 < 0.00000000000000022
##
## <none>
## - OLDCLAIM
## - BLUEBOOK
## - SEX         .
## - CLM_FREQ    **
## - REVOKED     ***
## - PARENT1     ***
## - CAR_AGE     ***
## - TIF         ***
## - TRAVTIME    ***
## - KIDSDRIV    ***
## - CAR_TYPE    ***
## - INCOME      ***
## - MSTATUS     ***
## - CAR_USE     ***
```

```
## - MVR_PTS    ***
## - URBANICITY ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Note that although bluebook was not dropped from the model it is not significant, thus we will drop from the final model. After dropping bluebook from the model, the variables of sex, old claim, and kids at home were no longer significant, thus these too were also dropped from the model.

```
mlr_final <- lm(TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS +
                  TRAVTIME + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED +
                  MVR_PTS + CAR_AGE + URBANICITY,
                data = df_insur_train[,-c(1, 26:27)])
```

```
summary(mlr_final)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS +
##     TRAVTIME + CAR_USE + TIF + CAR_TYPE + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY, data = df_insur_train[, -c(1,
##     26:27)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5741  -1683   -777    290 103795
##
## Coefficients:
##                          Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)           703.351694  250.323383   2.810             0.004969 **
## KIDSDRIV              701.963073  161.961661   4.334     0.00001480957868 ***
## INCOME                 -0.005970    0.001248  -4.783     0.00000176067979 ***
## PARENT11              584.661486  177.821928   3.288             0.001014 **
## MSTATUS1             -614.146456  119.579422  -5.136     0.00000028732293 ***
## TRAVTIME               12.713678    3.218079   3.951     0.00007857951135 ***
## CAR_USEPrivate       -857.847821  125.538896  -6.833     0.00000000000889 ***
## TIF                   -47.508100   12.168895  -3.904     0.00009535561187 ***
## CAR_TYPEPanel Truck   469.752522  227.438458   2.065             0.038916 *
## CAR_TYPEPickup        323.920408  164.886983   1.964             0.049506 *
## CAR_TYPESports Car    753.587800  181.708008   4.147     0.00003399156608 ***
## CAR_TYPESUV           502.708330  137.871553   3.646             0.000268 ***
## CAR_TYPEVan           609.482122  200.369238   3.042             0.002359 **
## CLM_FREQ              110.334260   48.809748   2.260             0.023817 *
## REVOKED1              468.774739  154.891289   3.026             0.002482 **
## MVR_PTS               179.996365   25.768921   6.985     0.00000000000307 ***
## CAR_AGE               -36.155650   10.045073  -3.599             0.000321 ***
## URBANICITYUrban      1530.115333  135.651793  11.280 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4550 on 8143 degrees of freedom
## Multiple R-squared:  0.06635,    Adjusted R-squared:  0.0644
## F-statistic: 34.04 on 17 and 8143 DF,  p-value: < 0.00000000000000022
```

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

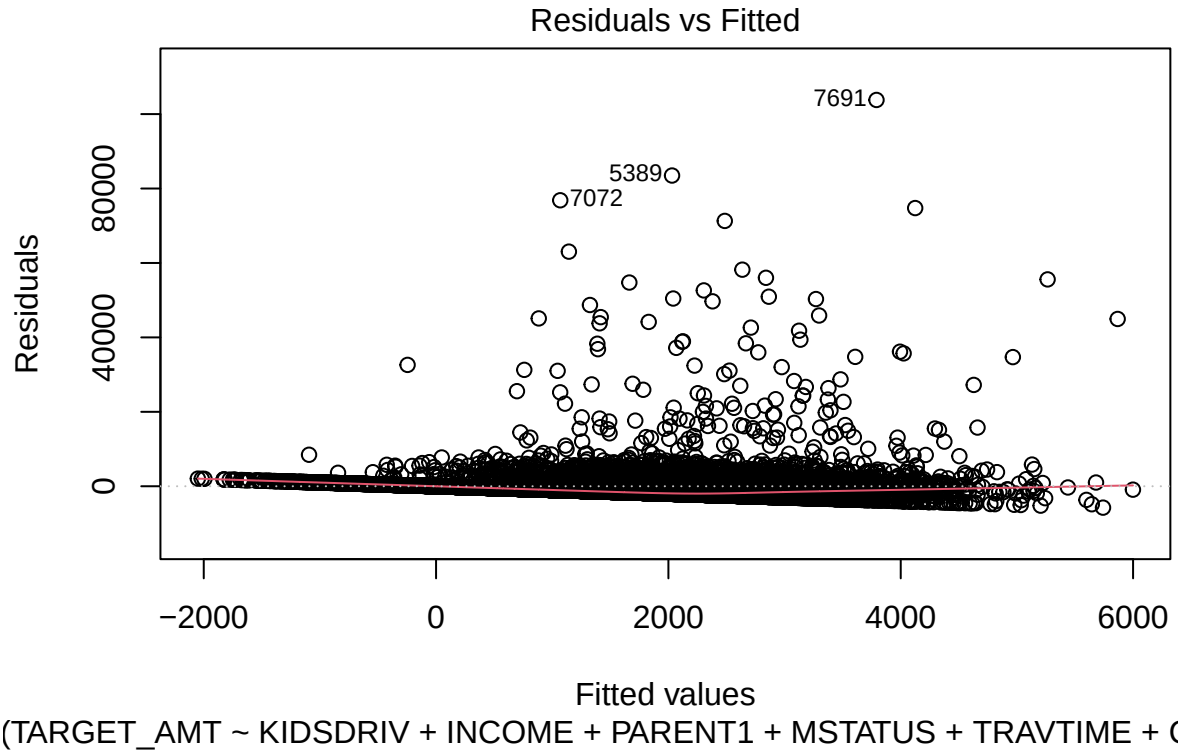
- Kids driving
- Being a parent(vs not being a a parent)
- Having a longer travel time
- Having a car type other than minivan(when compared to minivan)
- Having an increased claims frequency
- Having a revoked license
- Residing in an urban environment
- Having higher points on drivers license
-

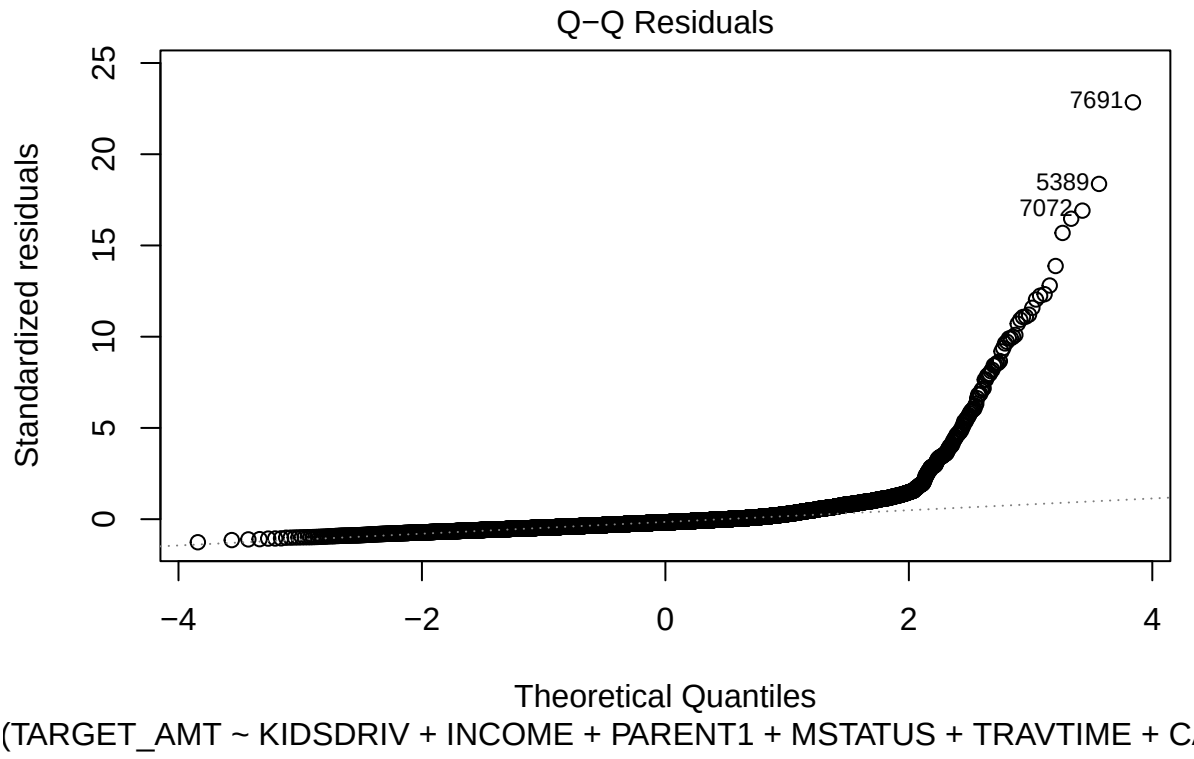
The variable that negatively impact the average cost of having crash are the following:

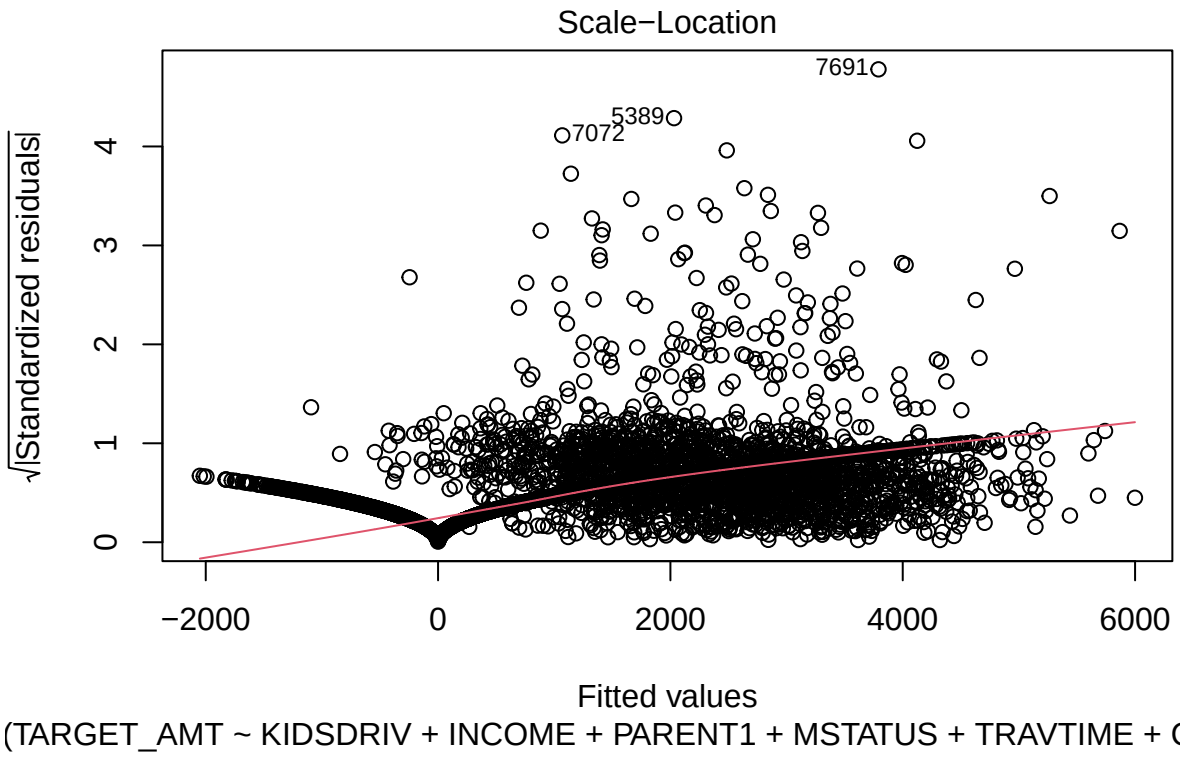
- Having a higher income
- Being married
- Using the car for private as opposed to commercial use
- Having a longer tenure as insurance client
- Having an older car

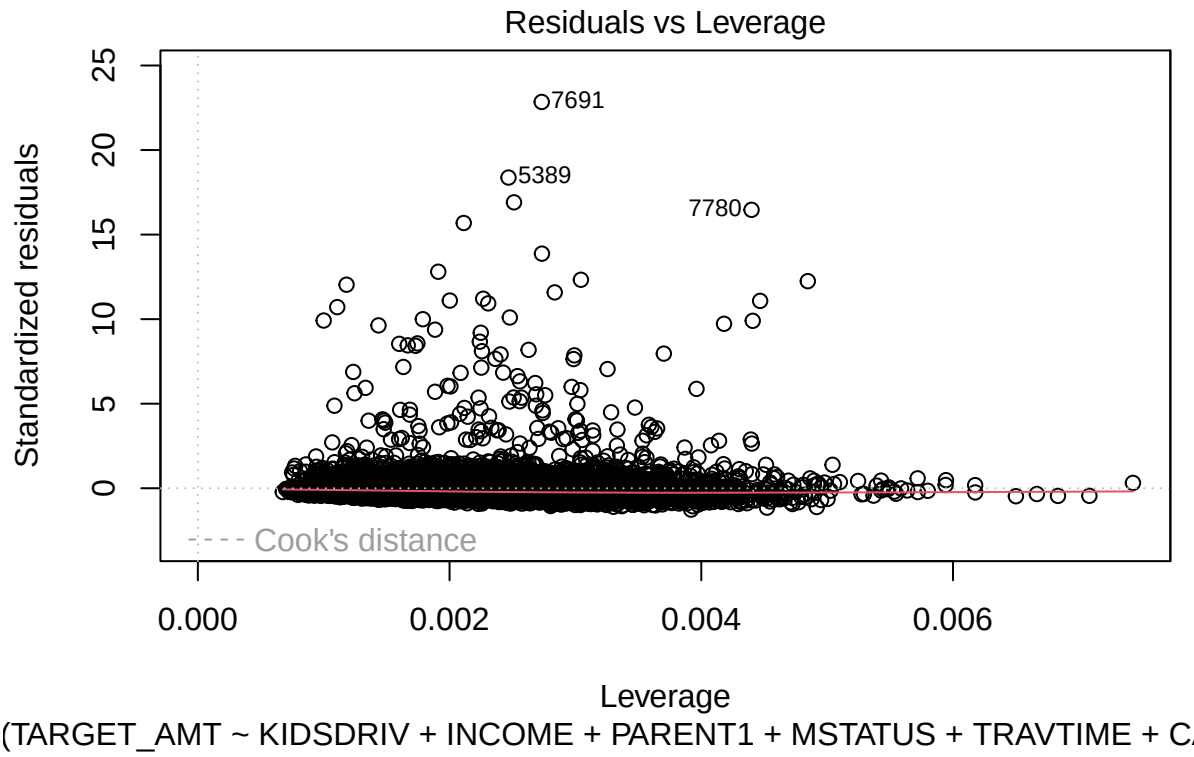
2.0.3 TEST MODEL ASSUMPTIONS

```
plot(mlr_final)
```




Q−Q Residuals
Standardized residuals
0
5
10
15
20
25
7691
5389
7072
−4
−2
0
2
4
Theoretical Quantiles
(TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS + TRAVTIME + CAR_


Scale−Location
7691
5389
7072
√|Standardized residuals|
0
1
2
3
4
−2000
0
2000
4000
6000
Fitted values
(TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS + TRAVTIME + CAR_



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 there is an approximate normal distribution of the residuals in the final model. 3. Equality of Variances - the third plot that there are some 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, but they do not violate cooks distance.

- Thus we can trust the results of the final model.

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

```
## [1] 4545.013
```

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

```
## [1] 0.06440114
```

- The RMSE for this model suggest an average deviation in the predicted claim amount from the true claim amount of $4,545. 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 only explain 6.4% of the total variation in the claim amount.

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

```
coef(mlr_final)
```

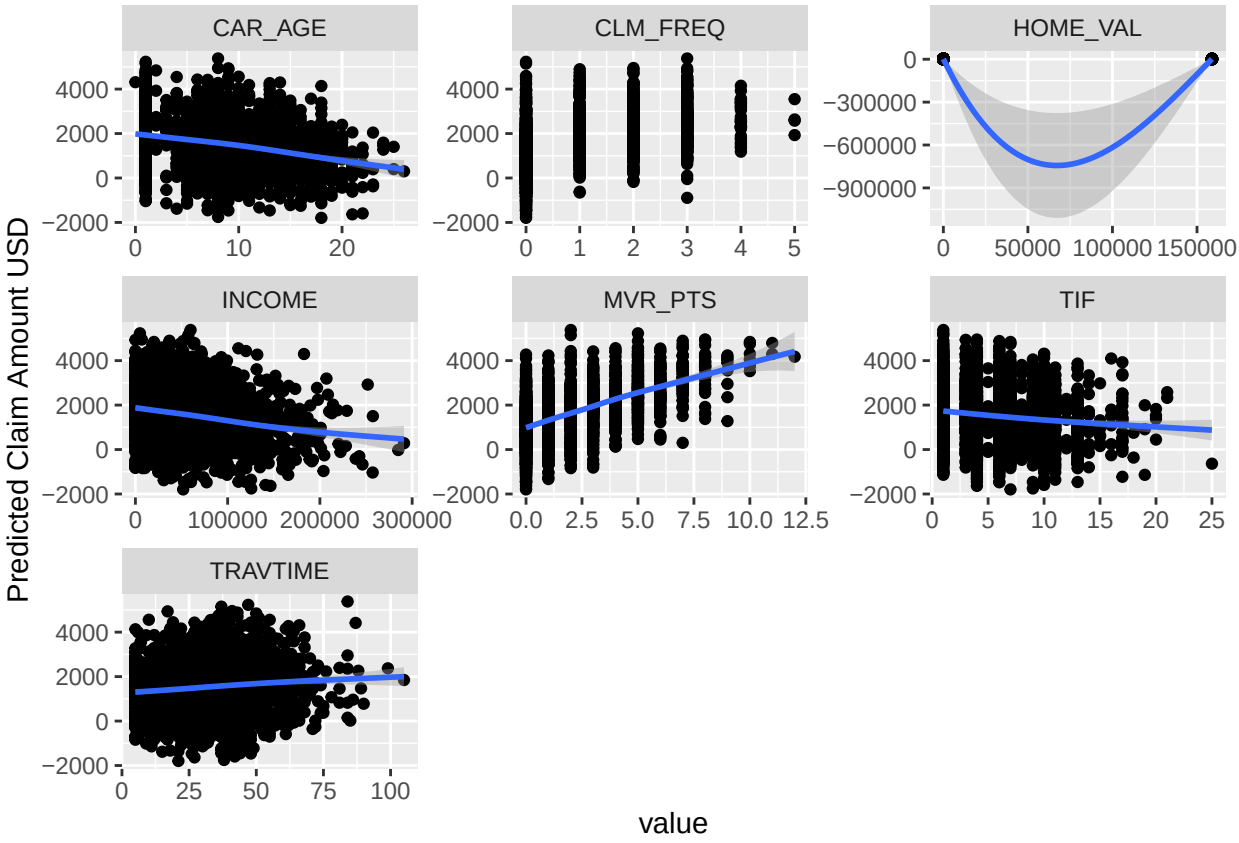
2.0.4.1 PREDICTING AMOUNT OF CLAIM FOR CAR CRASHES WITH THE EVALUATIONS DATASET

```
##         (Intercept)            KIDSDRIV              INCOME            PARENT11
##       703.351694436       701.963073253        -0.005969975       584.661485838
##            MSTATUS1            TRAVTIME      CAR_USEPrivate                 TIF
##      -614.146456160        12.713677618      -857.847820656       -47.508099735
## CAR_TYPEPanel Truck      CAR_TYPEPickup  CAR_TYPESports Car         CAR_TYPESUV
##       469.752522433       323.920408295       753.587799612       502.708329951
##          CAR_TYPEVan            CLM_FREQ            REVOKED1             MVR_PTS
##       609.482121905       110.334260103       468.774738749       179.996364559
##             CAR_AGE     URBANICITYUrban
##       -36.155650023      1530.115333365
```

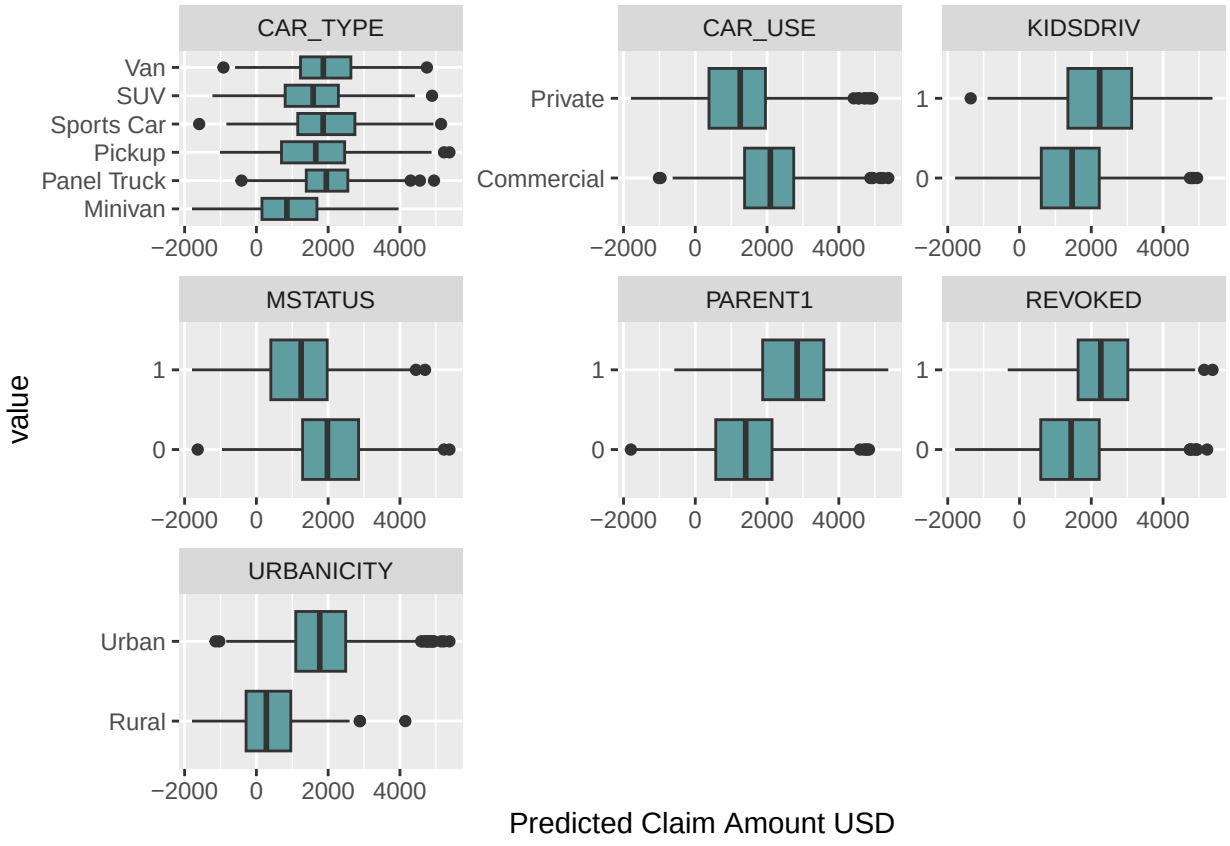
```
df_insur_eval %>%
  select(mlr_pred, INCOME, HOME_VAL, TRAVTIME, TIF, CLM_FREQ, MVR_PTS,
         CAR_AGE) %>%
  mutate_if(is.character, as.numeric) %>%
  pivot_longer(-mlr_pred, names_to = "key", values_to = "value") %>%
  ggplot(aes(x = value, y = mlr_pred)) +
  geom_point() +
  geom_smooth() +
  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.
```




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


CAR_TYPE
Van
SUV
Sports Car
Pickup
Panel Truck
Minivan
CAR_USE
Private
Commercial
KIDSDRIV
MSTATUS
PARENT1
REVOKED
URBANICITY
Urban
Rural
−2000
0
2000
4000
value
Predicted Claim Amount USD