DATA621_HW4

Avery Davidowitz, Gabriel Santos, John Ledesma, Josh Iden, Mathew Katz, Tyler Brown

2023-04-29

```
df <- read.csv("https://raw.githubusercontent.com/GabrielSantos33/DATA621_G2/main/DATA621_HW4/insurance_
evaluation <- read.csv("https://raw.githubusercontent.com/GabrielSantos33/DATA621_G2/main/DATA621_HW4/in
strip_dollars <- function(x){
  x <- as.character(x)
  x <- gsub(",", "", x)
  x <- gsub("\\$", "", x)
  as.numeric(x)
}
```

Objetive

The goal is to train a logistic regression classifier to predict whether a person was in a car accident, and to predict the insurance claim cost of the crash.

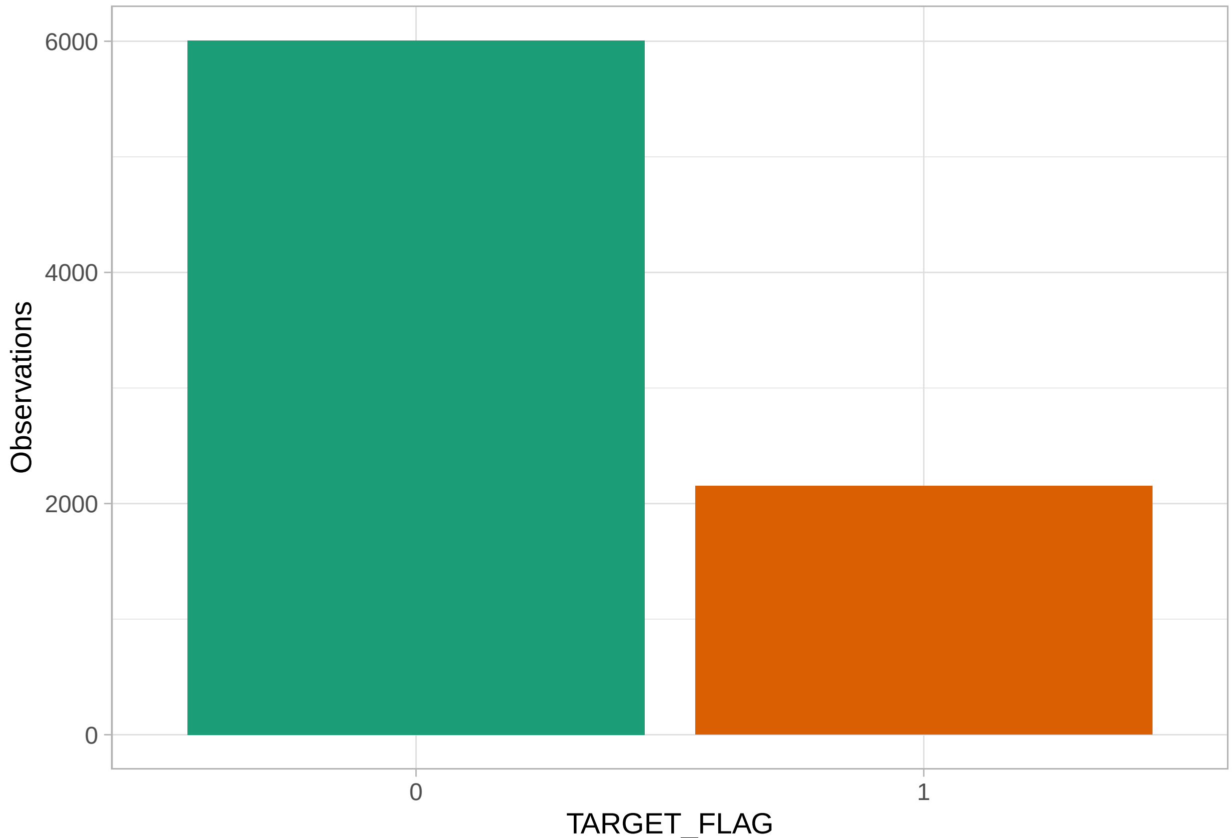
Introduction

We have a dataset with 8161 records representing customers of an auto insurance company. Each record has two response variables.

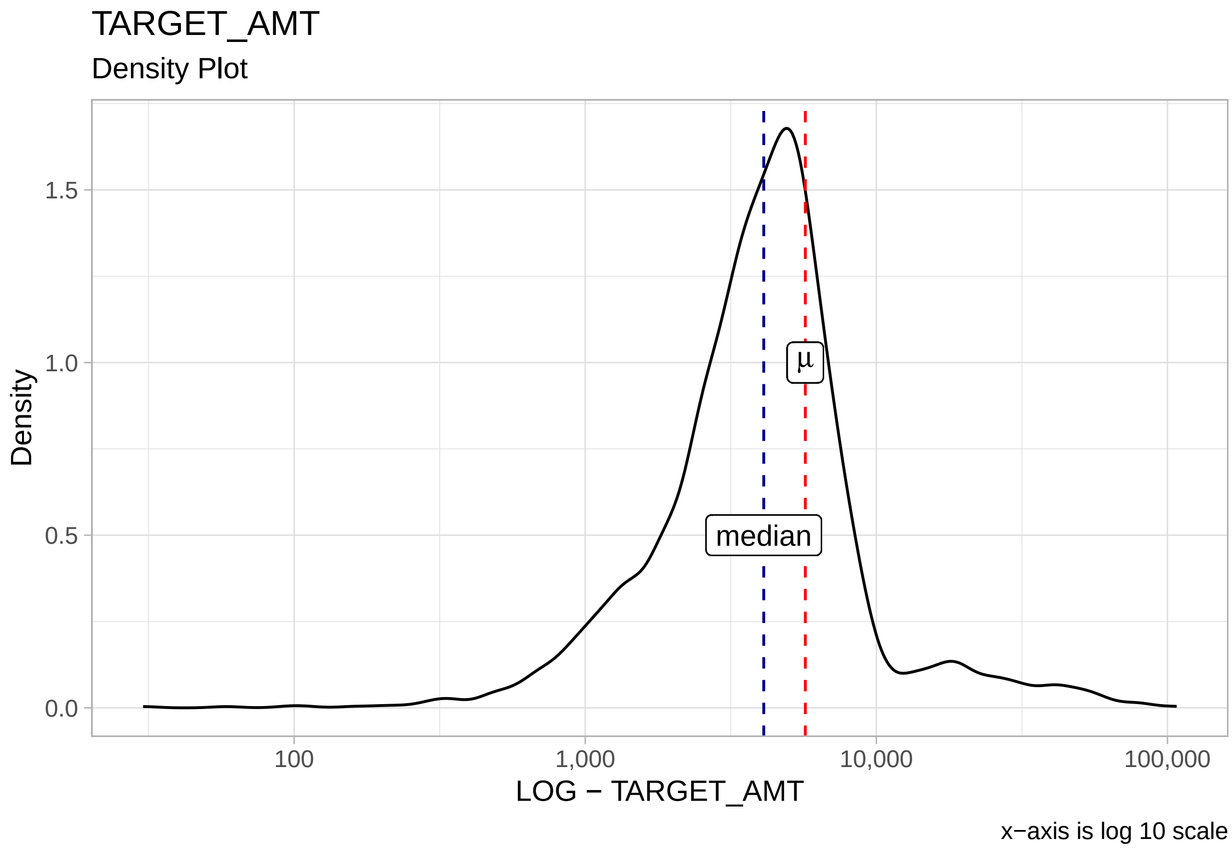
The first response variable is 'TARGET_FLAG' which represents whether a person had an accident (1) or did not have an accident (0). The second response variable is 'TARGET_AMT'.

This value is zero if the person did not crash their car. But if they crashed their car, this number will be a value greater than zero.

TARGET FLAG:



TARGET AMT:



From the graph we can see that the distribution of the 'TARGET_AMT' variable is skewed to the right. We thought we could apply the LOG transformation.

Data

Preparation & Exploration

Summary statistics for the data:

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", "", "$114,986", "$125,301~
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No~
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "$0"~
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", "Yes",~
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", "M"~
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High School", "~
```

```
## $ JOB        <chr> "Professional", "z_Blue Collar", "Clerical", "z_Blue Colla~
## $ TRAVTIME   <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE    <chr> "Private", "Commercial", "Private", "Private", "Private", ~
## $ BLUEBOOK   <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000", "$17~
## $ TIF        <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE   <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV", "Sports~
## $ RED_CAR    <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes", "no", ~
## $ OLDCLAIM   <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0", "$~
## $ CLM_FREQ   <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED    <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "N~
## $ MVR_PTS    <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE    <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly Urba~

##      INDEX       TARGET_FLAG      TARGET_AMT       KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ          INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Length:8161
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   Class :character
##  Median :45.00   Median :0.0000   Median :11.0   Mode  :character
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0
##  NA's   :6                        NA's   :454
##    PARENT1            HOME_VAL           MSTATUS              SEX
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   EDUCATION             JOB               TRAVTIME        CAR_USE
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
##                                        3rd Qu.: 44.00
##                                        Max.   :142.00
##
##    BLUEBOOK             TIF           CAR_TYPE           RED_CAR
##  Length:8161        Min.   : 1.000   Length:8161        Length:8161
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 5.351
##                     3rd Qu.: 7.000
##                     Max.   :25.000
##
```

	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ
	Min. : 1	Min. :0.0000	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000	Min. : 0.
	1st Qu.: 2559	1st Qu.:0.0000	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000	1st Qu.:
	Median : 5133	Median :0.0000	Median : 0	Median :0.0000	Median :45.00	Median :0.0000	Median :
	Mean : 5152	Mean :0.2638	Mean : 1504	Mean :0.1711	Mean :44.79	Mean :0.7212	Mean :10
	3rd Qu.: 7745	3rd Qu.:1.0000	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:1
	Max. :10302	Max. :1.0000	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000	Max. :23
	NA	NA	NA	NA	NA's :6	NA	NA's :45

	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ
	Min. : 1	0:6008	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000	Min. : 0.
	1st Qu.: 2559	1:2153	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000	1st Qu.:
	Median : 5133	NA	Median : 0	Median :0.0000	Median :45.00	Median :0.0000	Median :
	Mean : 5152	NA	Mean : 1504	Mean :0.1711	Mean :44.79	Mean :0.7212	Mean :10
	3rd Qu.: 7745	NA	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:1
	Max. :10302	NA	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000	Max. :23
	NA	NA	NA	NA	NA's :6	NA	NA's :45

```
##    OLDCLAIM           CLM_FREQ       REVOKED            MVR_PTS
##  Length:8161        Min.   :0.0000   Length:8161        Min.   : 0.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :0.0000   Mode  :character   Median : 1.000
##                     Mean   :0.7986                      Mean   : 1.696
##                     3rd Qu.:2.0000                      3rd Qu.: 3.000
##                     Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE        URBANICITY
##  Min.   :-3.000   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
##  Mean   : 8.328
##  3rd Qu.:12.000
##  Max.   :28.000
##  NA's   :510
```

To better observe the data we will use Kable package:

We can see that there are missing data. There is also data that has outliers, for example negative values in the variable 'CAR_AGE'

There are values that are represented in currency, we must change them to numerical values.

There are also some invalid data that will be changed to NAs.

Summary of the data with the corrected data:

Fix Missing Values

There are 1714, or 21% of the observations missing variables.

We will fill in the missing data with the median value.

```
## # A tibble: 8,161 x 26
##    INDEX TARGET_FLAG TARGET~1 KIDSD~2   AGE HOMEK~3   YOJ INCOME PARENT1 HOME_~4
```

	x
INDEX	0
TARGET_FLAG	0
TARGET_AMT	0
KIDSDRIV	0
AGE	6
HOMEKIDS	0
YOJ	454
INCOME	445
PARENT1	0
HOME_VAL	464
MSTATUS	0
SEX	0
EDUCATION	0
JOB	0
TRAVTIME	0
CAR_USE	0
BLUEBOOK	0
TIF	0
CAR_TYPE	0
RED_CAR	0
OLDCLAIM	0
CLM_FREQ	0
REVOKED	0
MVR_PTS	0
CAR_AGE	511
URBANICITY	0

```
##    <int> <fct>          <dbl>   <int> <int>   <int> <int>  <dbl> <chr>     <dbl>
##  1     1 0                  0       0    60       0    11  67349 No            0
##  2     2 0                  0       0    43       0    11  91449 No       257252
##  3     4 0                  0       0    35       1    10  16039 No       124191
##  4     5 0                  0       0    51       0    14  54028 No       306251
##  5     6 0                  0       0    50       0    11 114986 No       243925
##  6     7 1               2946       0    34       1    12 125301 Yes           0
##  7     8 0                  0       0    54       0    11  18755 No       161160
##  8    11 1               4021       1    37       2    11 107961 No       333680
##  9    12 1               2501       0    34       0    10  62978 No            0
## 10    13 0                  0       0    50       0     7 106952 No            0
## # ... with 8,151 more rows, 16 more variables: MSTATUS <chr>, SEX <chr>,
## #   EDUCATION <chr>, JOB <chr>, TRAVTIME <int>, CAR_USE <chr>, BLUEBOOK <dbl>,
## #   TIF <int>, CAR_TYPE <chr>, RED_CAR <chr>, OLDCLAIM <dbl>, CLM_FREQ <int>,
## #   REVOKED <chr>, MVR_PTS <int>, CAR_AGE <dbl>, URBANICITY <chr>, and
## #   abbreviated variable names 1: TARGET_AMT, 2: KIDSDRIV, 3: HOMEKIDS,
## #   4: HOME_VAL
```

Feature Creation

For 'INCOME' and HOME_VAL" we will apply log transformation. We create an average claim amount. We will identify outliers for "TARGET_ATM".

Function to add features:

Creating Data Sets (Training/Test)

For Classifier Model

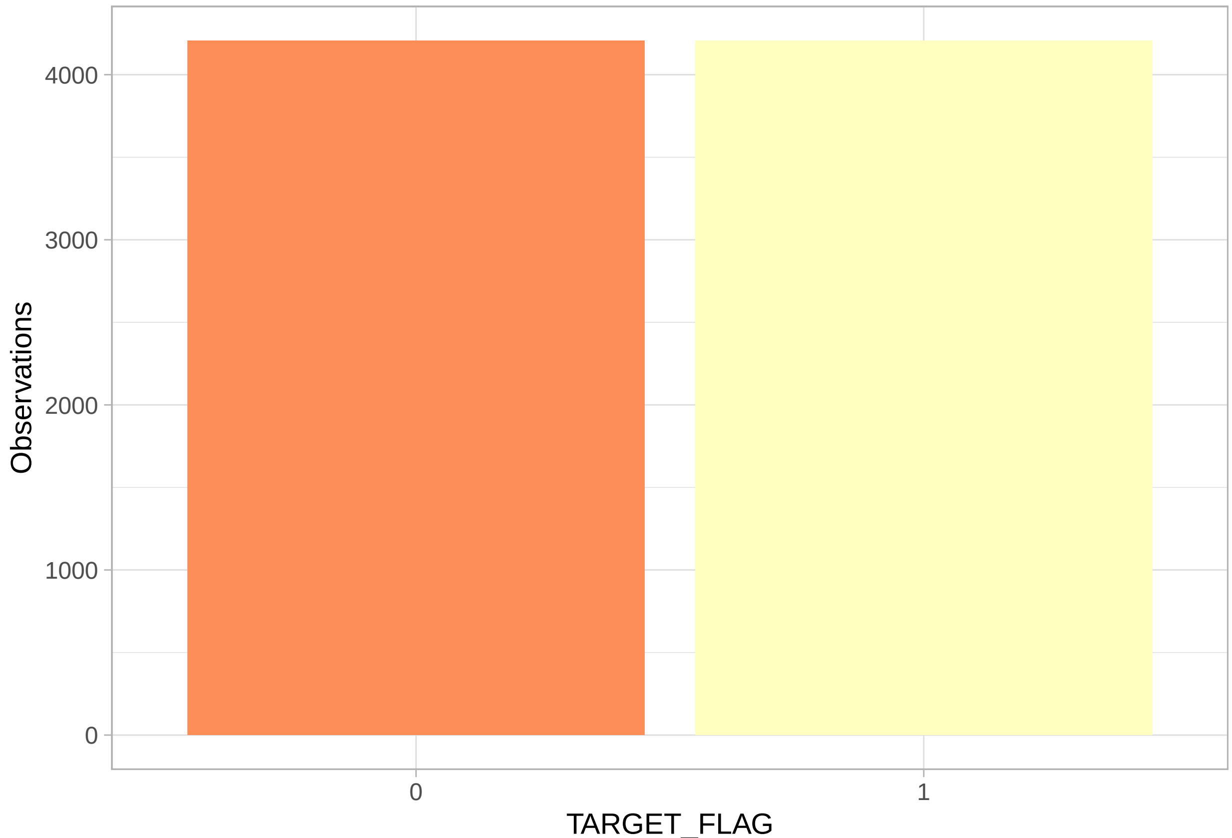
We will divide the data set into two groups, one for training and another for the test, 70% and 30% respectively.

```
## # A tibble: 5,714 x 37
##    INDEX TARGET_FLAG TARGET~1 KIDSD~2   AGE HOMEK~3   YOJ INCOME PARENT1 HOME_~4
##    <int> <fct>          <dbl>   <int> <int>   <int> <int>  <dbl> <chr>     <dbl>
##  1     2 0                  0       0    43       0    11  91449 No       257252
##  2     4 0                  0       0    35       1    10  16039 No       124191
##  3     5 0                  0       0    51       0    14  54028 No       306251
##  4     8 0                  0       0    54       0    11  18755 No       161160
##  5    11 1               4021       1    37       2    11 107961 No       333680
##  6    12 1               2501       0    34       0    10  62978 No            0
##  7    13 0                  0       0    50       0     7 106952 No            0
##  8    14 1               6077       0    53       0    14  77100 No            0
##  9    15 0                  0       0    43       0     5  52642 No       209970
## 10    16 0                  0       0    55       0    11  59162 No       180232
## # ... with 5,704 more rows, 27 more variables: MSTATUS <chr>, SEX <chr>,
## #   EDUCATION <chr>, JOB <chr>, TRAVTIME <int>, CAR_USE <chr>, BLUEBOOK <dbl>,
## #   TIF <int>, CAR_TYPE <chr>, RED_CAR <chr>, OLDCLAIM <dbl>, CLM_FREQ <int>,
## #   REVOKED <chr>, MVR_PTS <int>, CAR_AGE <dbl>, LOG_INCOME <dbl>,
## #   LOG_HOME_VAL <dbl>, AVG_CLAIM <dbl>, PRIOR_ACCIDENT <fct>,
## #   COLLEGE_EDUCATED <fct>, URBAN_DRIVER <fct>, YOUNG_MALE <fct>, YOUNG <fct>,
## #   RED_SPORTS_CAR <fct>, HAS_KIDS <fct>, KID_DRIVERS <fct>, ...
```

We can see that there are 1508 records of 5714 records in the training data set that have been in an accident.

So that the classifier can correctly identify the records, we will oversample the records that have been involved in an accident.

The over sampled data frame has 8412 records.



We can see that the data is now balanced.

Linear Regression Model

There are 2153 accident records in the data set. We will divide the data set into two groups, one for training and another for the test, 70% and 30% respectively.

There are 1509 out of 2153 records in the training data set.

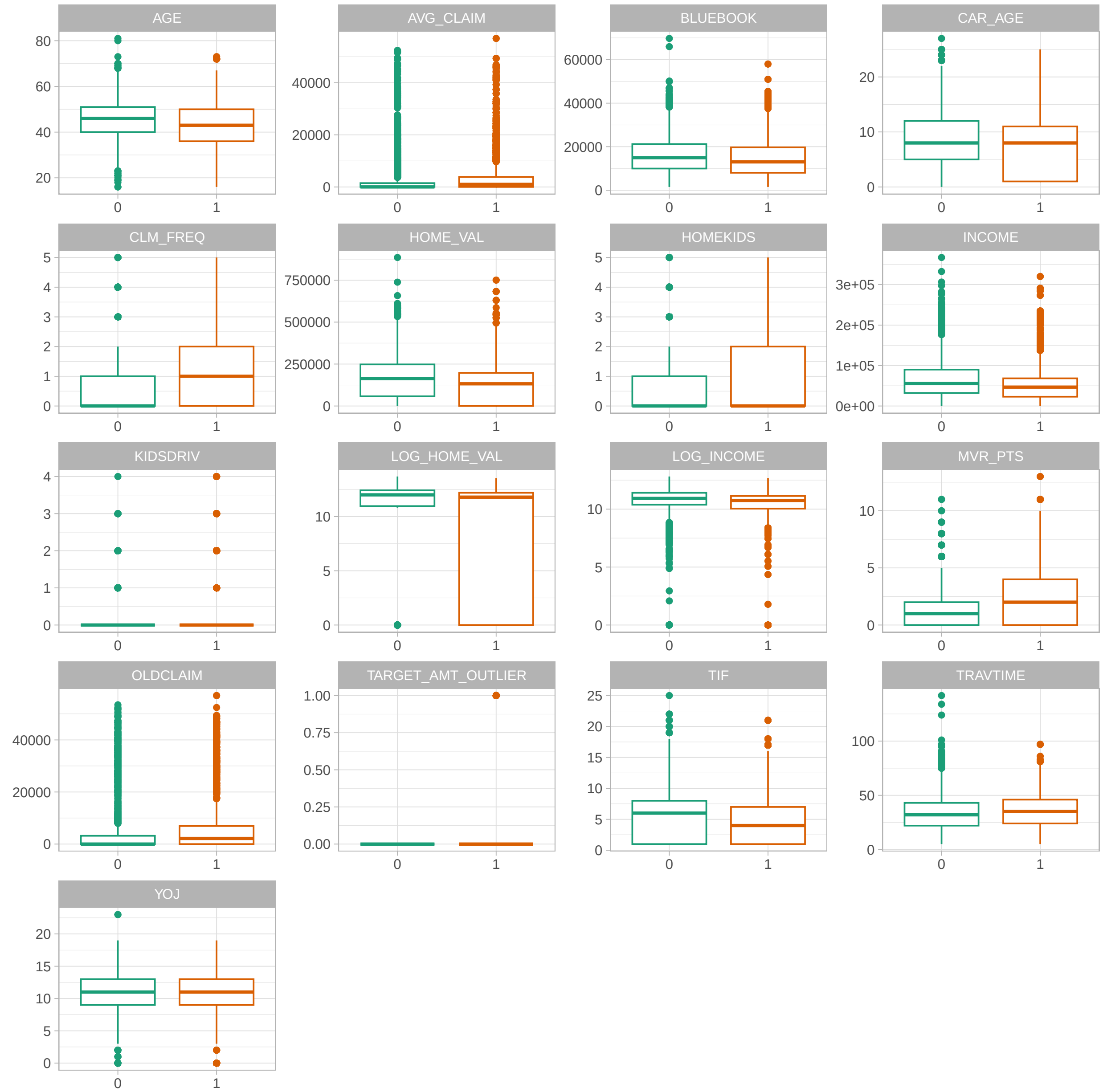
Exploratory Data Analysis

We are going to identify the variables that allow us to classify the data between those who have had an accident and those who have not.

We will identify the variables correlated with the claim amount and then use them as predictors for the linear regression model.

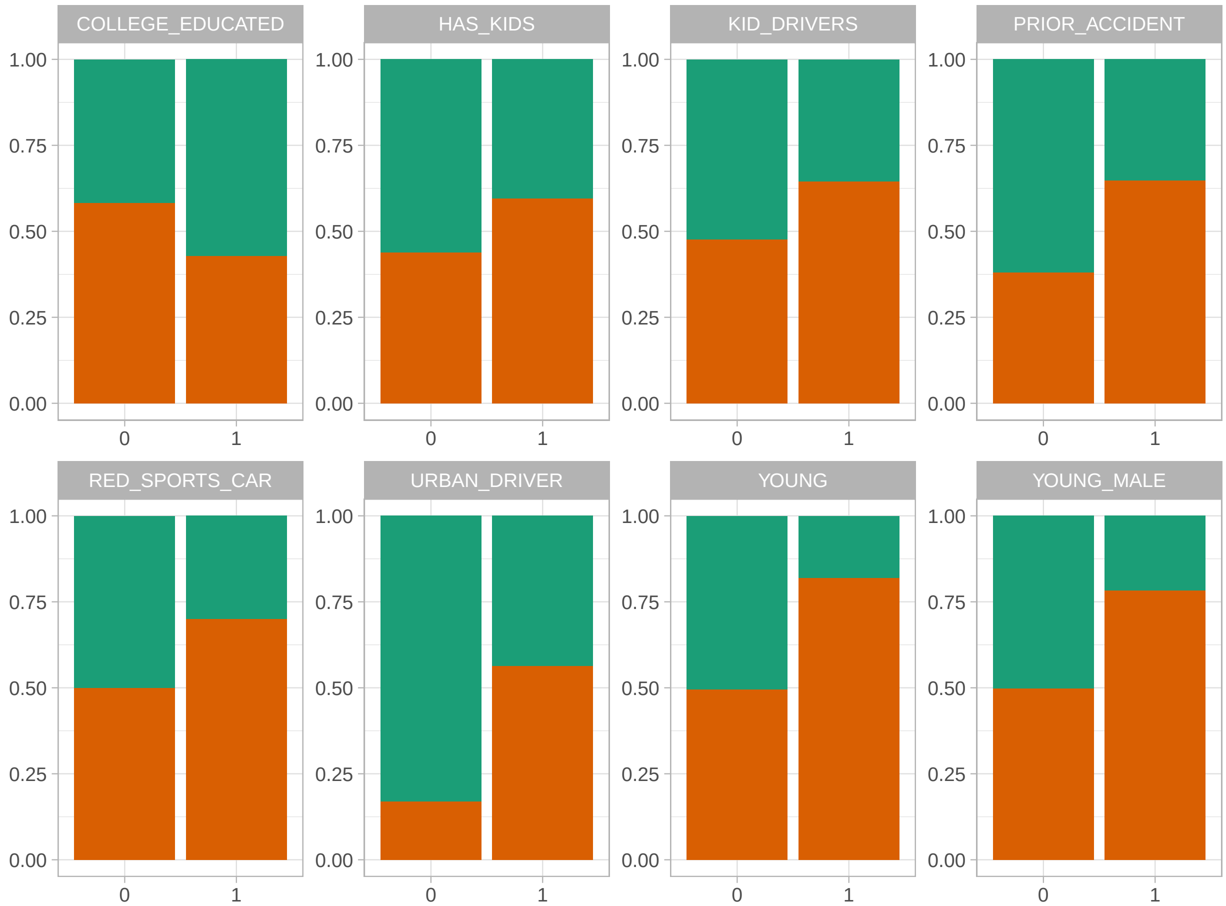
We will examine both training sets.

The oversampled classification data set:



The ‘CLM_FREQ’ variable seems to have a difference between the two groups. In general, it does not look different between the groups, whether a person had an accident (1) or did not have an accident (0).

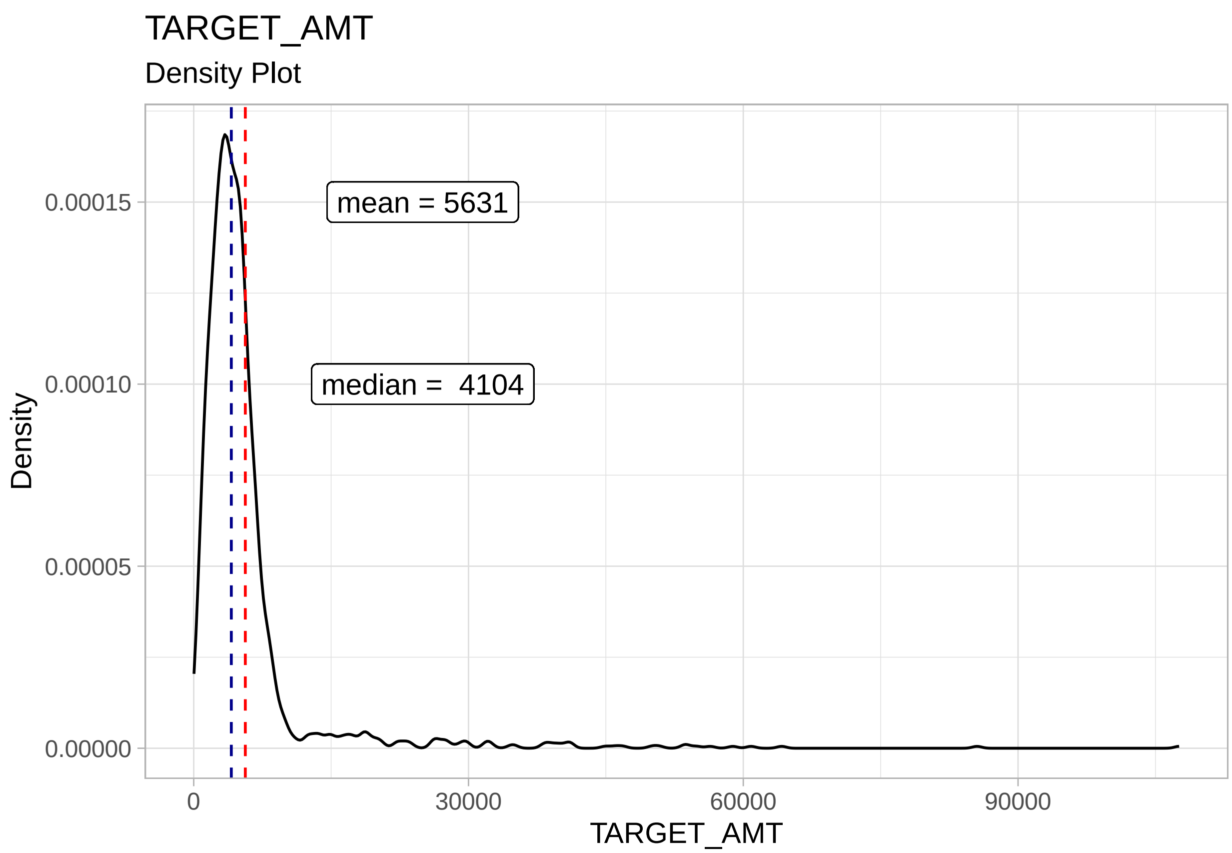
Categorically variables in the oversampled classification data set, the following graphs allow to identify if a variable can be used to distinguish those who have had an accident (orange) of those that are not (green)::



We can see that the 'URBAN_DRIVER' is more likely to have an accident, also the 'YOUNG' and those with a 'PRIOR_ACCIDENT'.

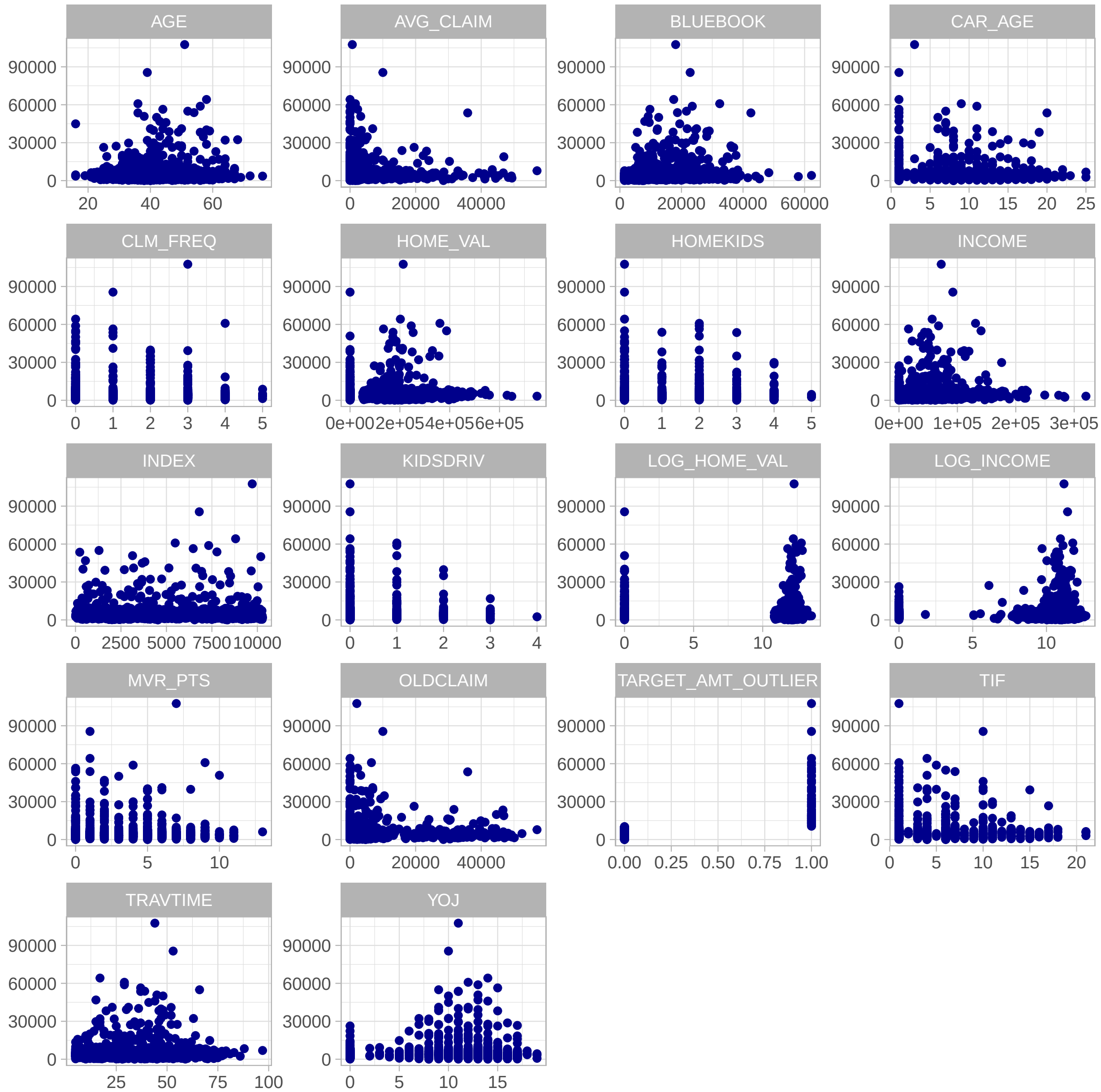
We will analyze the distribution of the claims of those who have had an accident:

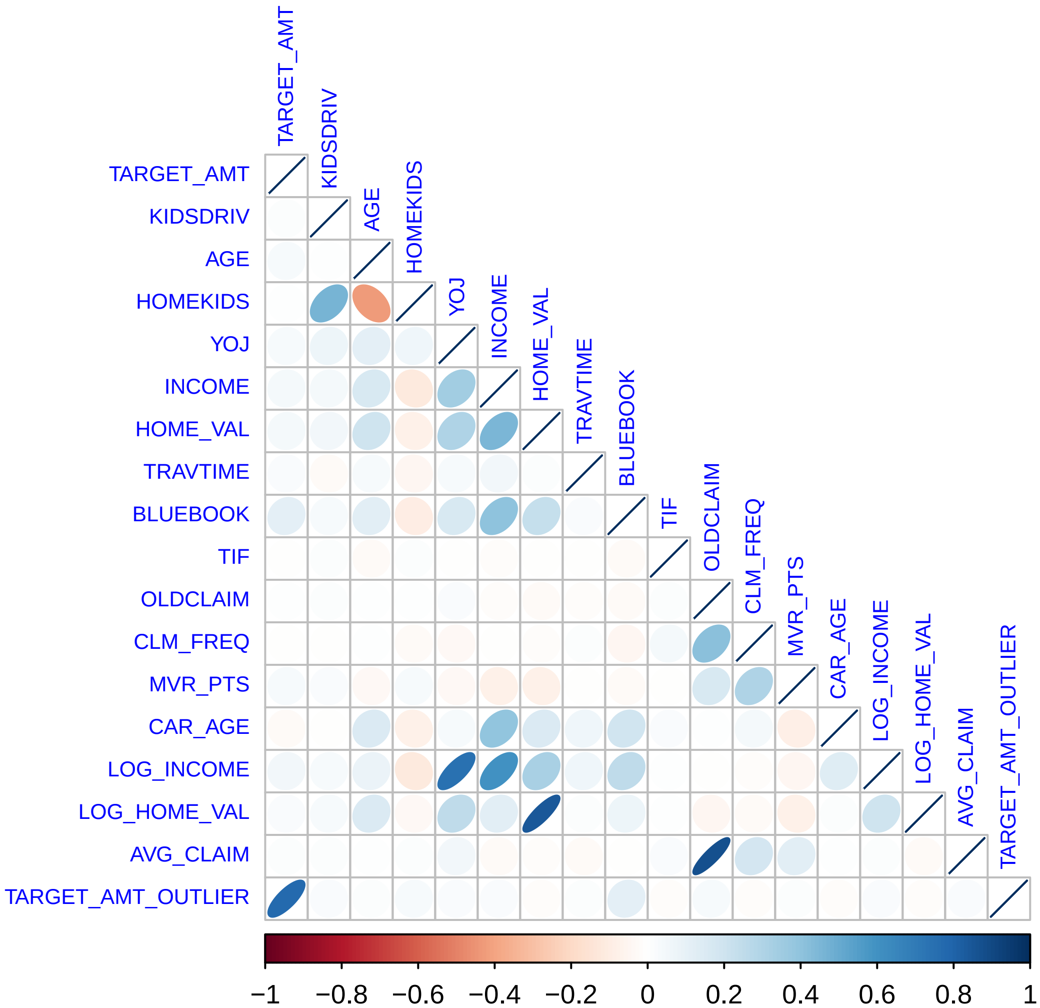
TARGET_AMT_OUTLIER	Mean	Median
No	4064.757	3917.00
Yes	26789.783	20279.68



The distribution is skewed to the left. The mean payout is 5631 dollars, and the median is $4104 dollars. The values are high, we can classify them as outliers.

We are going to make the correlation and dispersion graphs of the numerical variables, to identify the predictors of the amount of the claim:

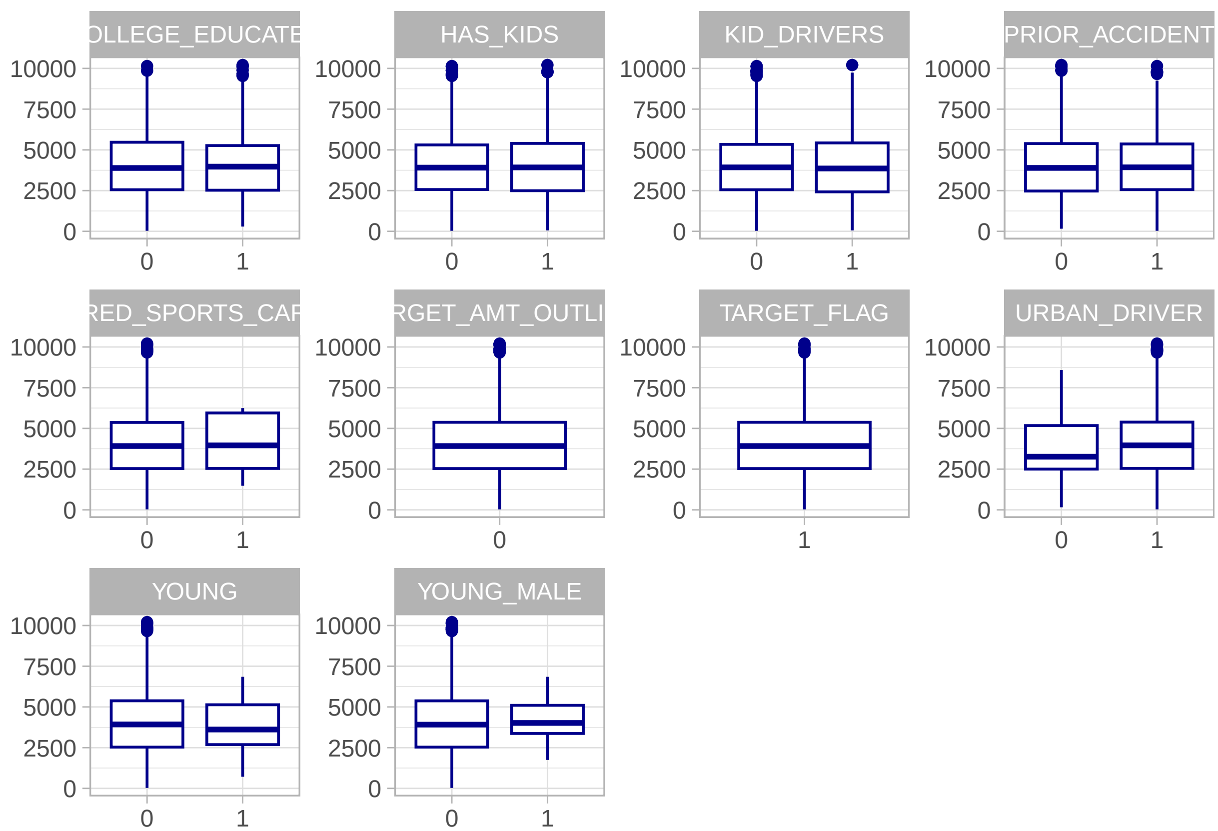

AGE
AVG_CLAIM
BLUEBOOK
CAR_AGE
CLM_FREQ
HOME_VAL
HOMEKIDS
INCOME
INDEX
KIDSDRIV
LOG_HOME_VAL
LOG_INCOME
MVR_PTS
OLDCLAIM
TARGET_AMT_OUTLIER
TIF
TRAVTIME
YOJ



In most of the predictors there is not a strong correlation with the amount of the claim. So far we can only choose the outliers that we identify.

Let's look at the categorical variables:

KIDSDRIV	0.0119333
AGE	0.0370895
HOMEKIDS	0.0015207
YOJ	0.0350769
INCOME	0.0489315
HOME_VAL	0.0407772
TRAVTIME	0.0210144
BLUEBOOK	0.1193868
TIF	-0.0095727
OLDCLAIM	0.0042991
CLM_FREQ	-0.0090870
MVR_PTS	0.0341212
CAR_AGE	-0.0261657
LOG_INCOME	0.0558976
LOG_HOME_VAL	0.0245686
AVG_CLAIM	0.0148793
TARGET_AMT_OUTLIER	0.7728035



The previous Boxplot confirms to us that there is no difference in the different groups for the amounts of the claims.

Analysis

According to the data exploration we have determined that there are no significant variables that allow us to differentiate the data and to be able to determine if there is a variable that affects the results. Possibly the accidents have been generated randomly and there is no variable that directly affects the number of accidents.

In order to predict the amount of the claim, we must carry out a deeper analysis because there are few variables that correlate with the amounts of the claims.

Classification Model

We will create predictive models and then analyze them.

For the classification models we use the test data.

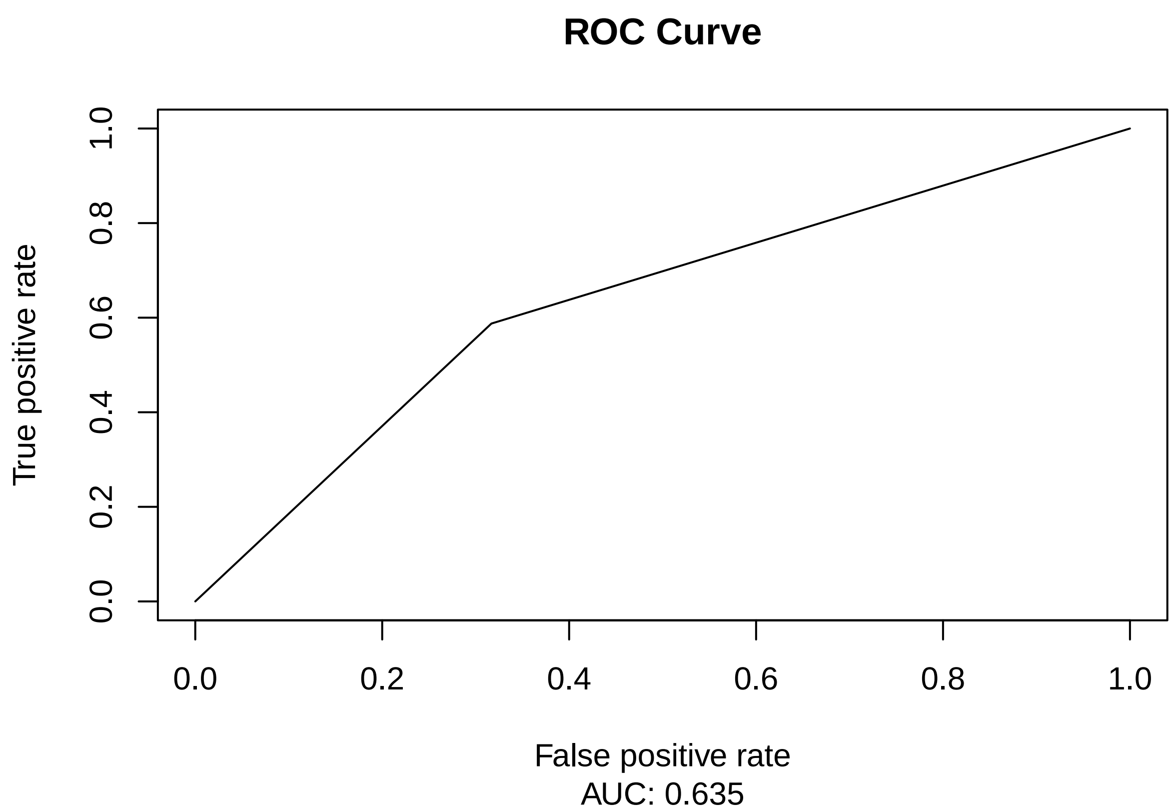
Baseline Model

We will create a simple model to serve as the baseline.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ PRIOR_ACCIDENT, family = binomial(link = "logit"),
##     data = over_sample_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.44507  -0.97775  -0.02311   0.93153   1.39114
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.48964    0.03023  -16.20   <2e-16 ***
## PRIOR_ACCIDENT1  1.09988    0.04558   24.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11662  on 8411  degrees of freedom
## Residual deviance: 11056  on 8410  degrees of freedom
## AIC: 11060
##
## Number of Fisher Scoring iterations: 4
```

```
## F1 = 0.4752351
## R2 = 0.05190635
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1231  266
```

```
##          1  571  379
##
##                Accuracy : 0.6579
##                  95% CI : (0.6388, 0.6768)
##     No Information Rate : 0.7364
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.235
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.5876
##             Specificity : 0.6831
##          Pos Pred Value : 0.3989
##          Neg Pred Value : 0.8223
##              Prevalence : 0.2636
##          Detection Rate : 0.1549
##    Detection Prevalence : 0.3882
##       Balanced Accuracy : 0.6354
##
##        'Positive' Class : 1
##
```



Drivers history is a representation of their future. Drivers who have been in an accident are more likely to have another accident. Drivers who haven't been in an accident probably won't be in one in the future.

Applying this model to the test data set indicates this simple model has a 65.7% accuracy rate.Correctly recognized 58.7% of the people with accidents and 67.3% of those without.

Let's see if other models can improve this precision.

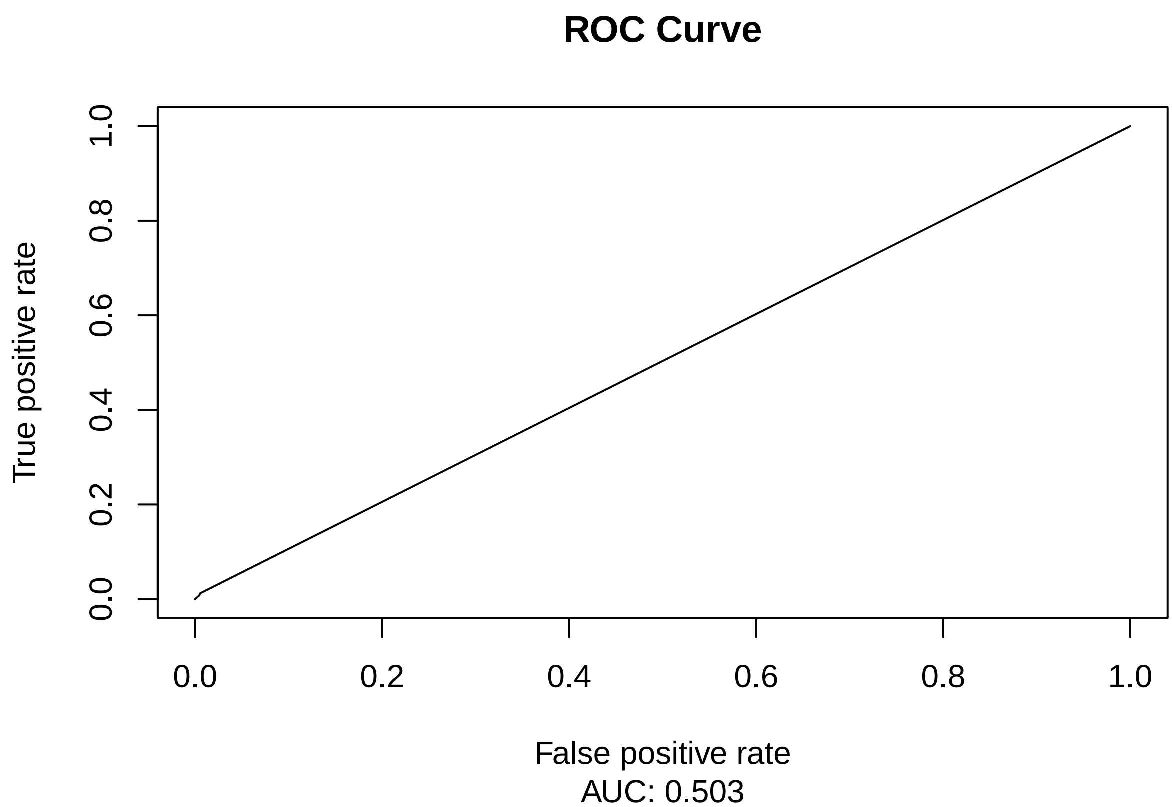
Risk Taker Model

To use this model, Let's assume that people who take more risks are more likely to have an accident. For this case we assume that young men take more risks.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ RED_SPORTS_CAR + YOUNG_MALE, family = binomial(link = "logit"),
##     data = over_sample_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7470  -1.1727  -0.2363   1.1821   1.1821
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.01106    0.02193  -0.504 0.613961
## RED_SPORTS_CAR1  0.85836    0.30938   2.774 0.005530 **
## YOUNG_MALE1      1.29200    0.35813   3.608 0.000309 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11662  on 8411  degrees of freedom
## Residual deviance: 11637  on 8409  degrees of freedom
## AIC: 11643
##
## Number of Fisher Scoring iterations: 4
```

```
## F1 = 0.02413273
## R2 = 0.002065113
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1792  637
##          1   10    8
##
##                Accuracy : 0.7356
##                  95% CI : (0.7176, 0.753)
##     No Information Rate : 0.7364
##     P-Value [Acc > NIR] : 0.5471
##
##                   Kappa : 0.01
##
##  Mcnemar's Test P-Value : <2e-16
##
```

```
##             Sensitivity : 0.012403
##             Specificity : 0.994451
##          Pos Pred Value : 0.444444
##          Neg Pred Value : 0.737752
##              Prevalence : 0.263588
##          Detection Rate : 0.003269
##    Detection Prevalence : 0.007356
##       Balanced Accuracy : 0.503427
##
##        'Positive' Class : 1
##
```



This model has a 73.5% accuracy rate. The model identified 99.4% of the people who didn't have an accident.The sensitivity of the model is 1.2%, this data means that it correctly identified the people who had an accident. We will not use this model.

Traditional Model

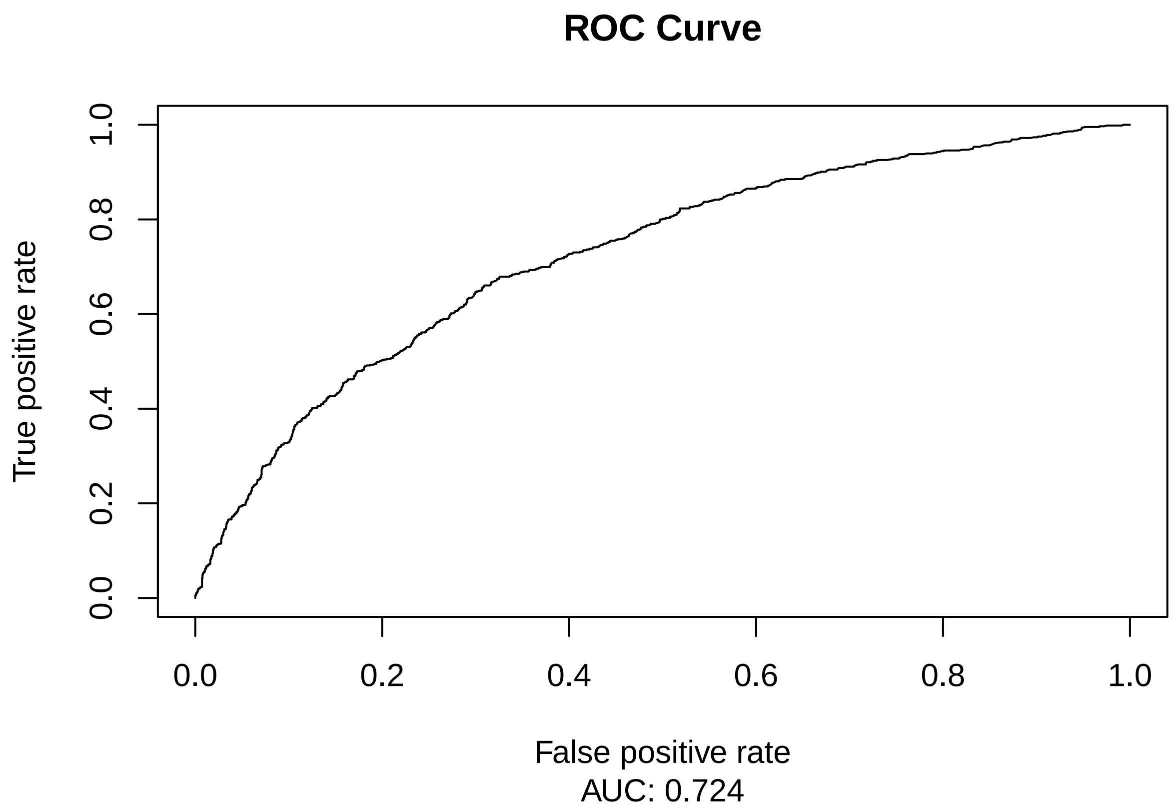
According to the analyzes that one can find of traffic accidents, there are some common predictors, for example, gender, age, accident history. We are going to use them in this model.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ YOUNG + MSTATUS + PRIOR_ACCIDENT +
##     SEX + REVOKED + MVR_PTS + TRAVTIME + CAR_USE, family = binomial(link = "logit"),
```

```
##     data = over_sample_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3240  -0.9978  -0.1906   1.0478   1.8809
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.956625   0.073816 -12.960  < 2e-16 ***
## YOUNG1           1.234430   0.263053   4.693 2.70e-06 ***
## MSTATUSz_No      0.524173   0.047624  11.006  < 2e-16 ***
## PRIOR_ACCIDENT1  0.804272   0.051675  15.564  < 2e-16 ***
## SEXz_F           0.246436   0.050102   4.919 8.71e-07 ***
## REVOKEDYes       0.885177   0.069552  12.727  < 2e-16 ***
## MVR_PTS          0.128330   0.011727  10.944  < 2e-16 ***
## TRAVTIME         0.007189   0.001487   4.835 1.33e-06 ***
## CAR_USEPrivate  -0.661273   0.050883 -12.996  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11662  on 8411  degrees of freedom
## Residual deviance: 10367  on 8403  degrees of freedom
## AIC: 10385
##
## Number of Fisher Scoring iterations: 4
```

```
## F1 = 0.5219001
## R2 = 0.1109659
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1249  222
##          1  553  423
##
##                Accuracy : 0.6833
##                  95% CI : (0.6644, 0.7017)
##     No Information Rate : 0.7364
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2996
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6558
##             Specificity : 0.6931
##          Pos Pred Value : 0.4334
##          Neg Pred Value : 0.8491
##              Prevalence : 0.2636
##          Detection Rate : 0.1729
##    Detection Prevalence : 0.3989
```

```
##       Balanced Accuracy : 0.6745
##
##        'Positive' Class : 1
##
```



This model has a 68.3% accuracy rate. It correcly identified 65.5% of the people with accidents and 69.3% of those without.
This model out preforms the baseline model.

Traditional Model with Cross-Validation

So far, the best result has been with the traditional model, we will try to improve the model with the cross-validation technique.

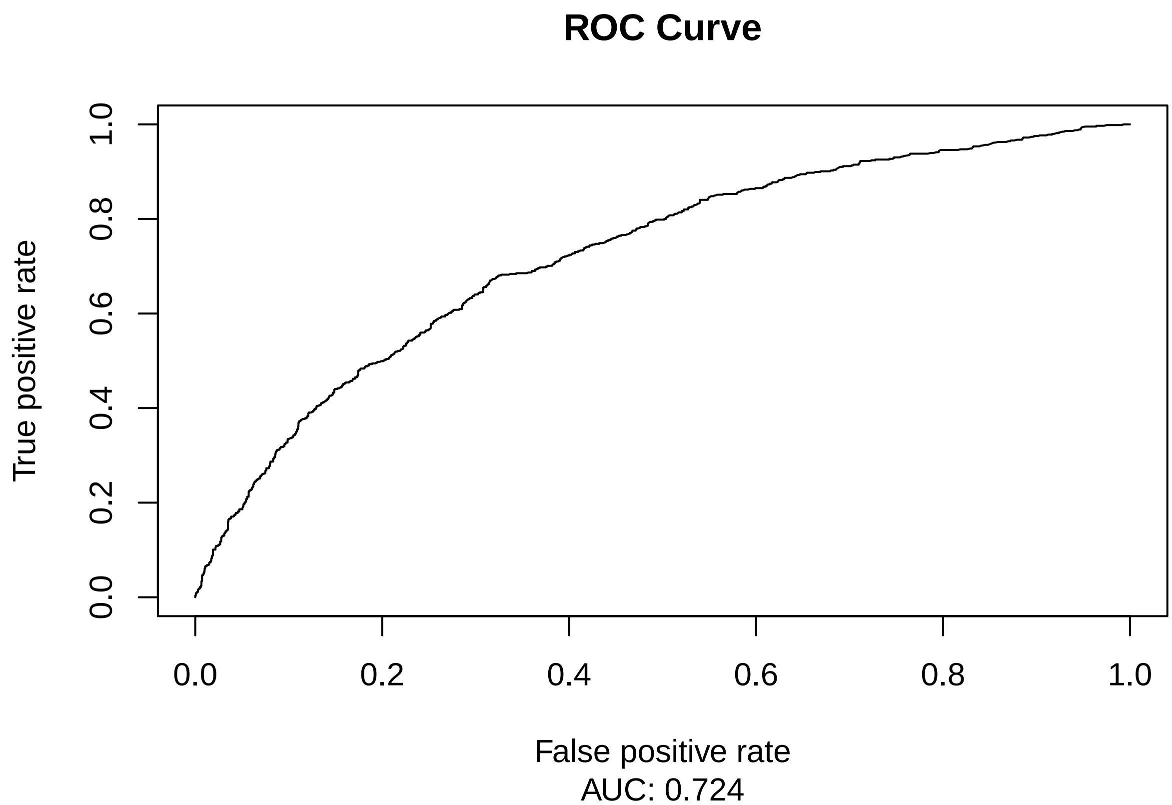
Let's use the original dataset and we are going to use 4 fold cross-validation:

```
## Generalized Linear Model
##
## 5714 samples
##    8 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (4 fold)
## Summary of sample sizes: 4285, 4286, 4285, 4286
## Addtional sampling using up-sampling
```

```
##
## Resampling results:
##
##   Accuracy   Kappa
##   0.6736045  0.2805572
```

Evaluating the model:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1256  231
##          1  546  414
##
##                Accuracy : 0.6825
##                  95% CI : (0.6636, 0.7009)
##     No Information Rate : 0.7364
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2929
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6419
##             Specificity : 0.6970
##          Pos Pred Value : 0.4313
##          Neg Pred Value : 0.8447
##              Prevalence : 0.2636
##          Detection Rate : 0.1692
##    Detection Prevalence : 0.3923
##       Balanced Accuracy : 0.6694
##
##        'Positive' Class : 1
##
```



This model has a 68.2% accuracy rate. It accurately recognized 64% of the people with accidents and 69.7% of those without. It is like the traditional model.

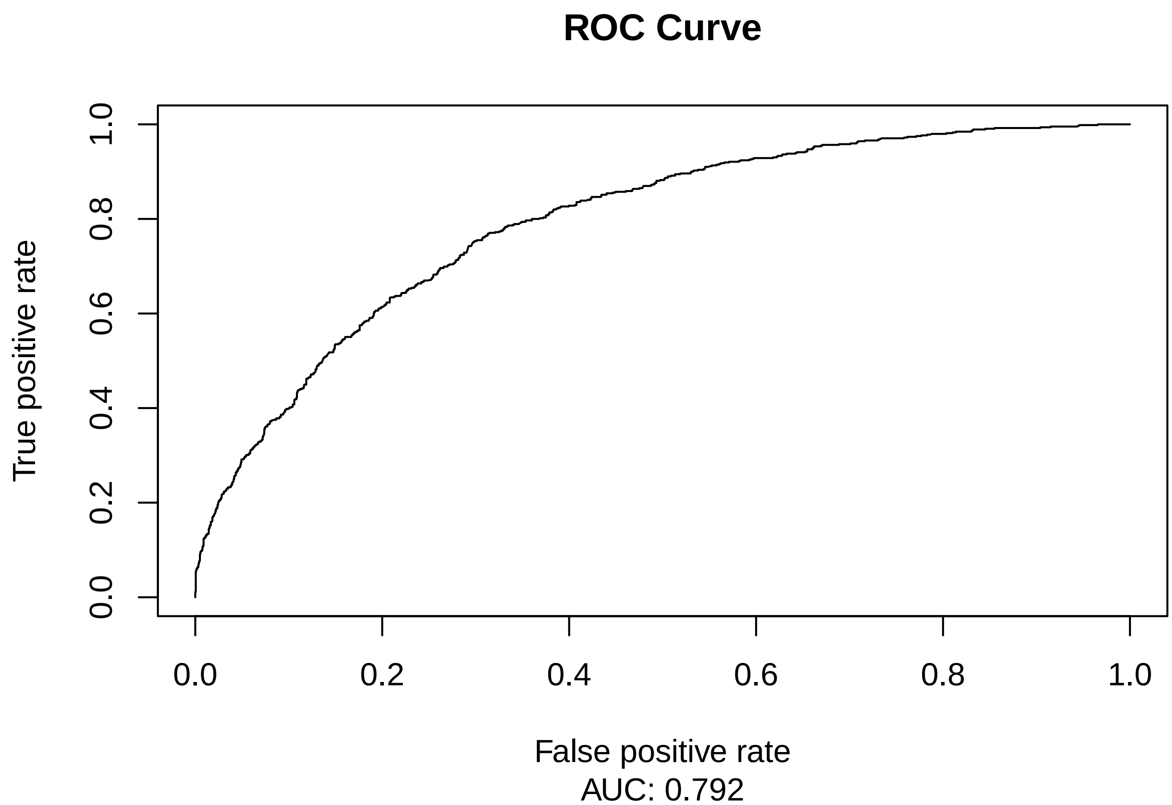
Alternate Traditional Model

This model is an alternate to the traditional model, taking into account other additional values.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ PRIOR_ACCIDENT + KID_DRIVERS + MSTATUS +
##     INCOME + SEX + CAR_USE + COLLEGE_EDUCATED + REVOKED + URBAN_DRIVER,
##     family = binomial(link = "logit"), data = over_sample_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.47073  -0.94301   0.03806   0.92022   2.61755
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.568e+00  9.475e-02 -16.545  < 2e-16 ***
## PRIOR_ACCIDENT1    7.189e-01  5.078e-02  14.157  < 2e-16 ***
## KID_DRIVERS1       8.775e-01  7.371e-02  11.904  < 2e-16 ***
## MSTATUSz_No        7.080e-01  5.111e-02  13.851  < 2e-16 ***
## INCOME            -7.563e-06  6.621e-07 -11.423  < 2e-16 ***
## SEXz_F             2.143e-01  5.316e-02   4.031 5.56e-05 ***
```

```
## CAR_USEPrivate    -8.110e-01  5.482e-02 -14.795  < 2e-16 ***
## COLLEGE_EDUCATED1 -5.569e-01  5.879e-02  -9.472  < 2e-16 ***
## REVOKEDYes         7.671e-01  7.283e-02  10.533  < 2e-16 ***
## URBAN_DRIVER1      2.097e+00  8.623e-02  24.315  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11661.5  on 8411  degrees of freedom
## Residual deviance:  9443.9  on 8402  degrees of freedom
## AIC: 9463.9
##
## Number of Fisher Scoring iterations: 4
```

```
## F1 = 0.5766801
## R2 = 0.1901654
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1208  143
##          1  594  502
##
##                Accuracy : 0.6988
##                  95% CI : (0.6802, 0.7169)
##     No Information Rate : 0.7364
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3664
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7783
##             Specificity : 0.6704
##          Pos Pred Value : 0.4580
##          Neg Pred Value : 0.8942
##              Prevalence : 0.2636
##          Detection Rate : 0.2051
##    Detection Prevalence : 0.4479
##       Balanced Accuracy : 0.7243
##
##        'Positive' Class : 1
##
```



This model has a 69.8% accuracy rate. It accurately recognized 77.8% of the people with accidents and 67% of those without.

Claims prediction

Baseline Model

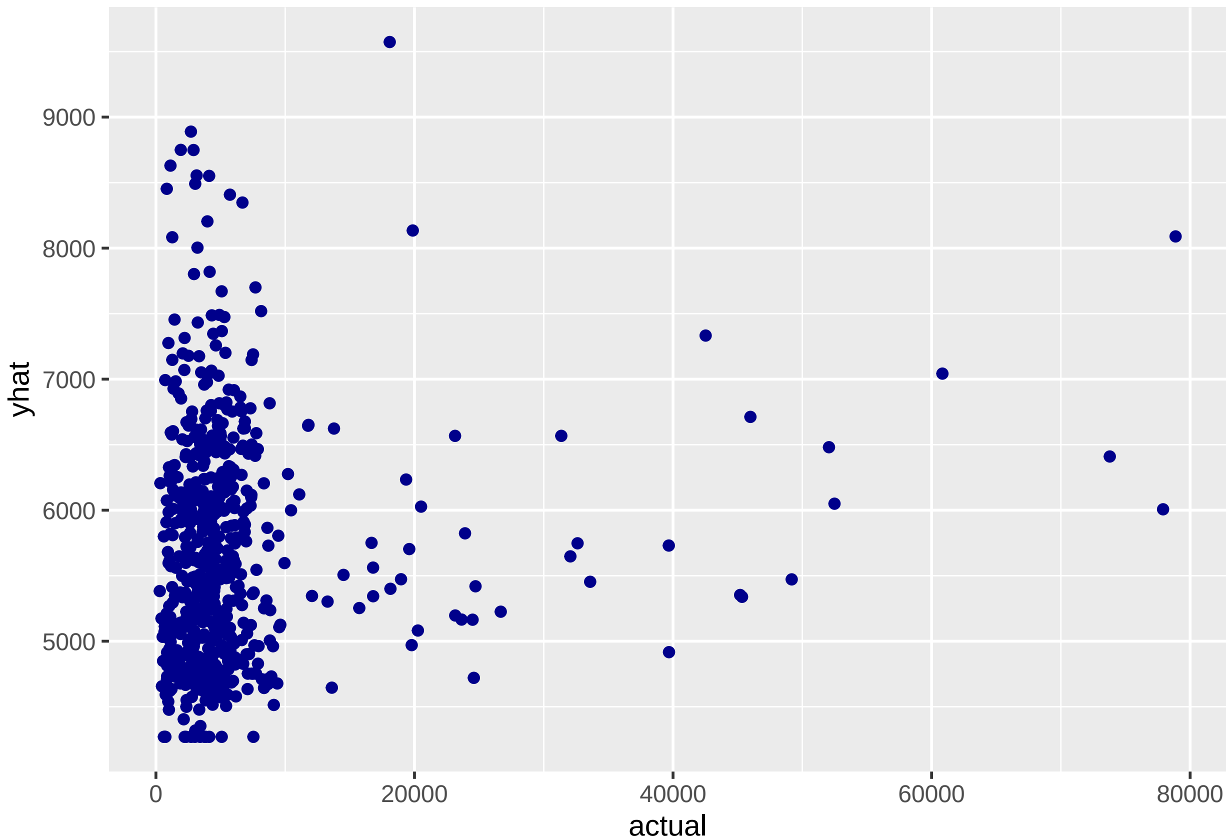
For this model we will assume that the claim amount is based on the value of the vehicle. More expensive vehicles should cost more to repair than less expensive vehicles.

```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK, data = amt_train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7571  -2979  -1507    382 101535
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.109e+03  3.776e+02  10.884  < 2e-16 ***
## BLUEBOOK    1.072e-01  2.296e-02   4.668 3.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 7401 on 1507 degrees of freedom
## Multiple R-squared:  0.01425,    Adjusted R-squared:  0.0136
## F-statistic: 21.79 on 1 and 1507 DF,  p-value: 3.312e-06
```

This predictor is statistically significant and positive.

Let's see how it performed on the test set:



Outlier Model

We are going to use the outliers that we determined earlier.

```
##
## Call:
## lm(formula = TARGET_AMT ~ TARGET_AMT_OUTLIER, data = amt_train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16196  -1646   -207   1331  80796
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4064.8      126.2   32.21   <2e-16 ***
## TARGET_AMT_OUTLIER  22725.0      480.7   47.27   <2e-16 ***
```

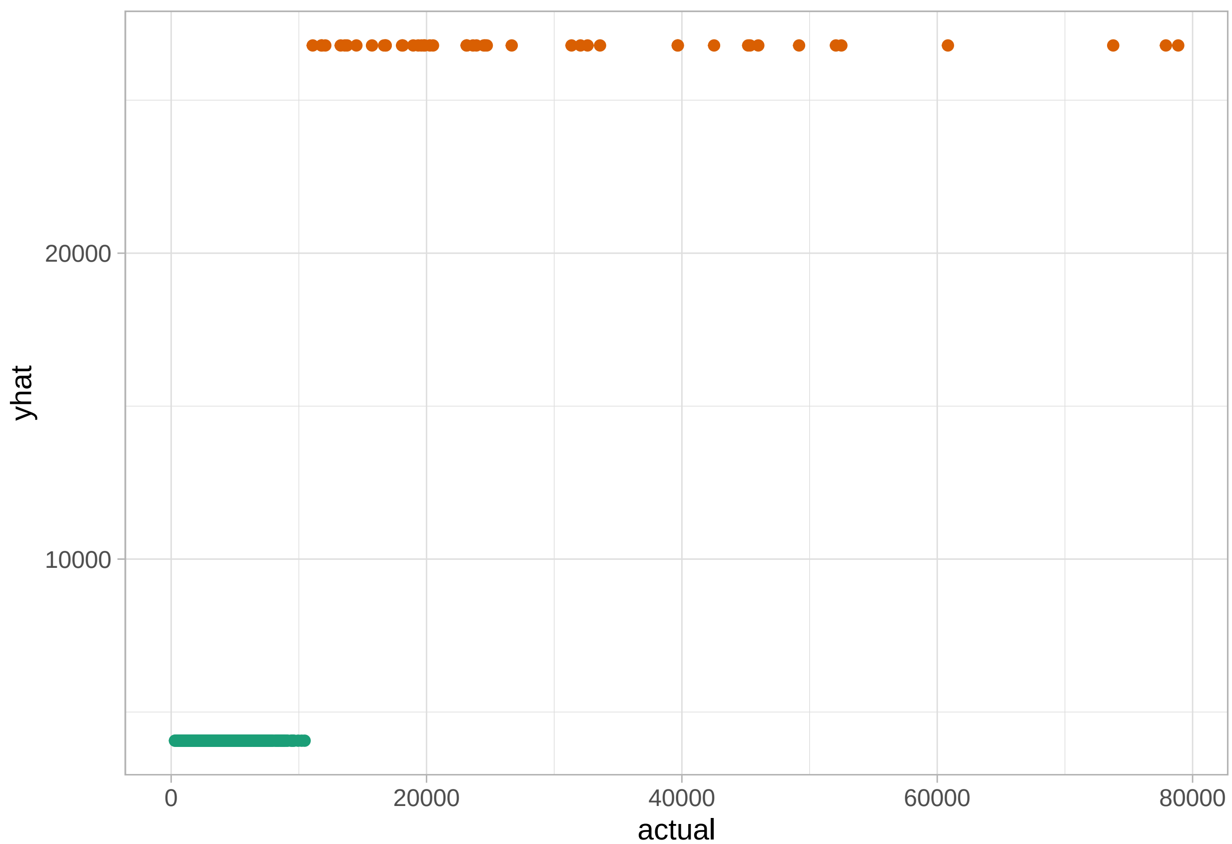
TARGET_AMT_OUTLIER	error	error %
0	35.10571	51.69964
1	-2991.52851	19.26778

TARGET_AMT_OUTLIER	error	error %
0	1601.587	107.96510
1	-23847.106	-73.79444

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4731 on 1507 degrees of freedom
## Multiple R-squared:  0.5972, Adjusted R-squared:  0.597
## F-statistic:  2235 on 1 and 1507 DF,  p-value: < 2.2e-16
```

This model appears to be incorrect because it predicts outcomes based on a predictor derived from an outcome. It has an adjusted R2 of 0.597.

Let's see how the model preforms on the test set:



The prediction is between 35 dollars for the lowest claims and $3300 for the large claims. The error on the model is about 51% of the estimate for the small claims and 19% for the large claims. Makes sense.

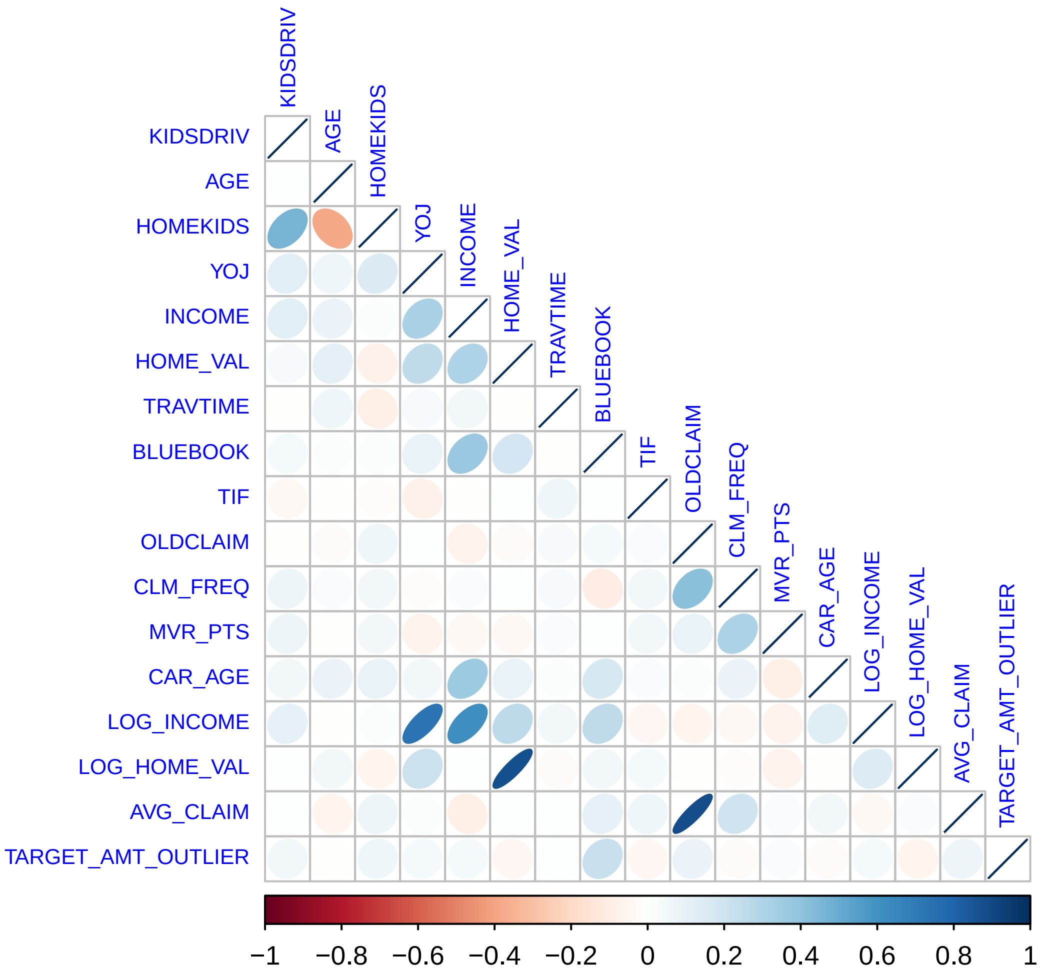
The table below offers the similar metrics:

We can see that they are outliers.

Let's create a classifier with a balanced data set:

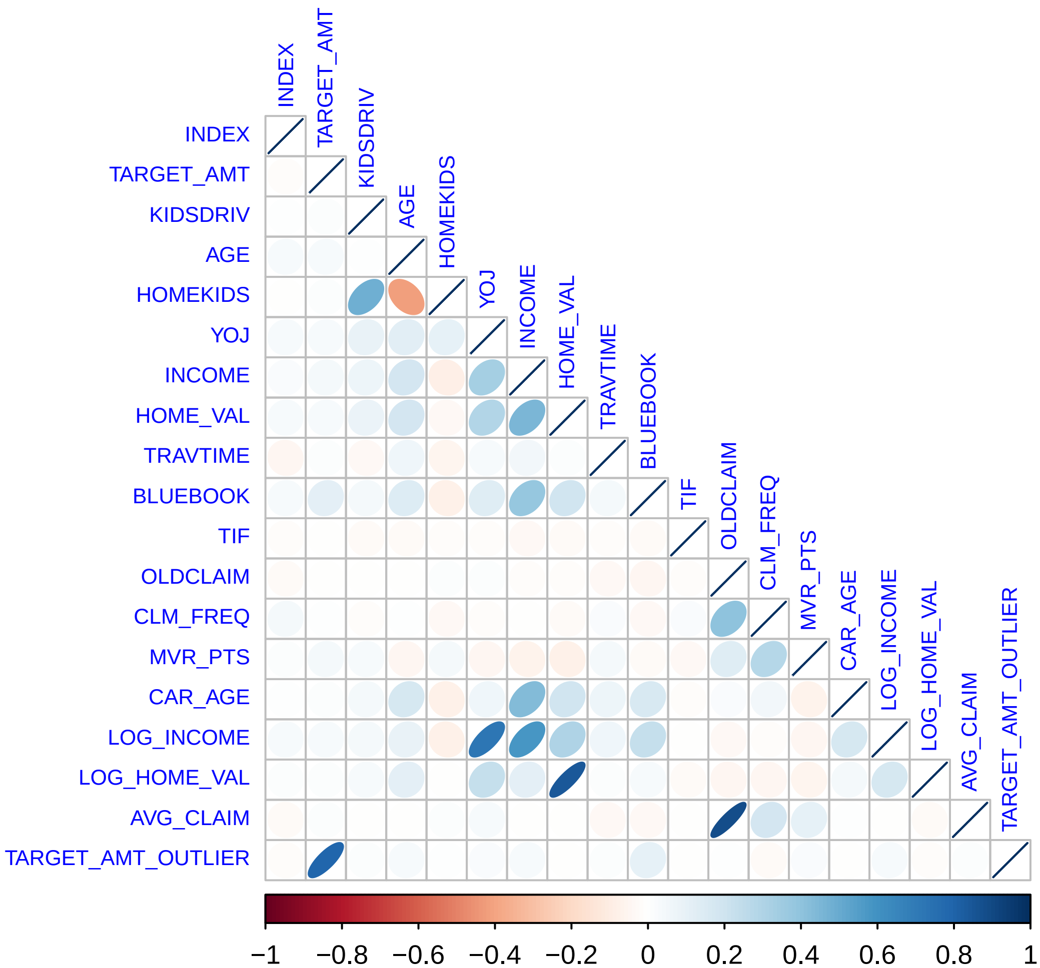
KIDSDRIV	0.0515117
AGE	-0.0015785
HOMEKIDS	0.0665627
YOJ	0.0477824
INCOME	0.0460100
HOME_VAL	-0.0431769
TRAVTIME	0.0036037
BLUEBOOK	0.2289962
TIF	-0.0417230
OLDCLAIM	0.0852083
CLM_FREQ	-0.0287347
MVR_PTS	0.0268186
CAR_AGE	-0.0220531
LOG_INCOME	0.0443220
LOG_HOME_VAL	-0.0567037
AVG_CLAIM	0.0729539

Let's make the correlation graphs:



Urban Model

Let's filter by URBAN_DRIVER:

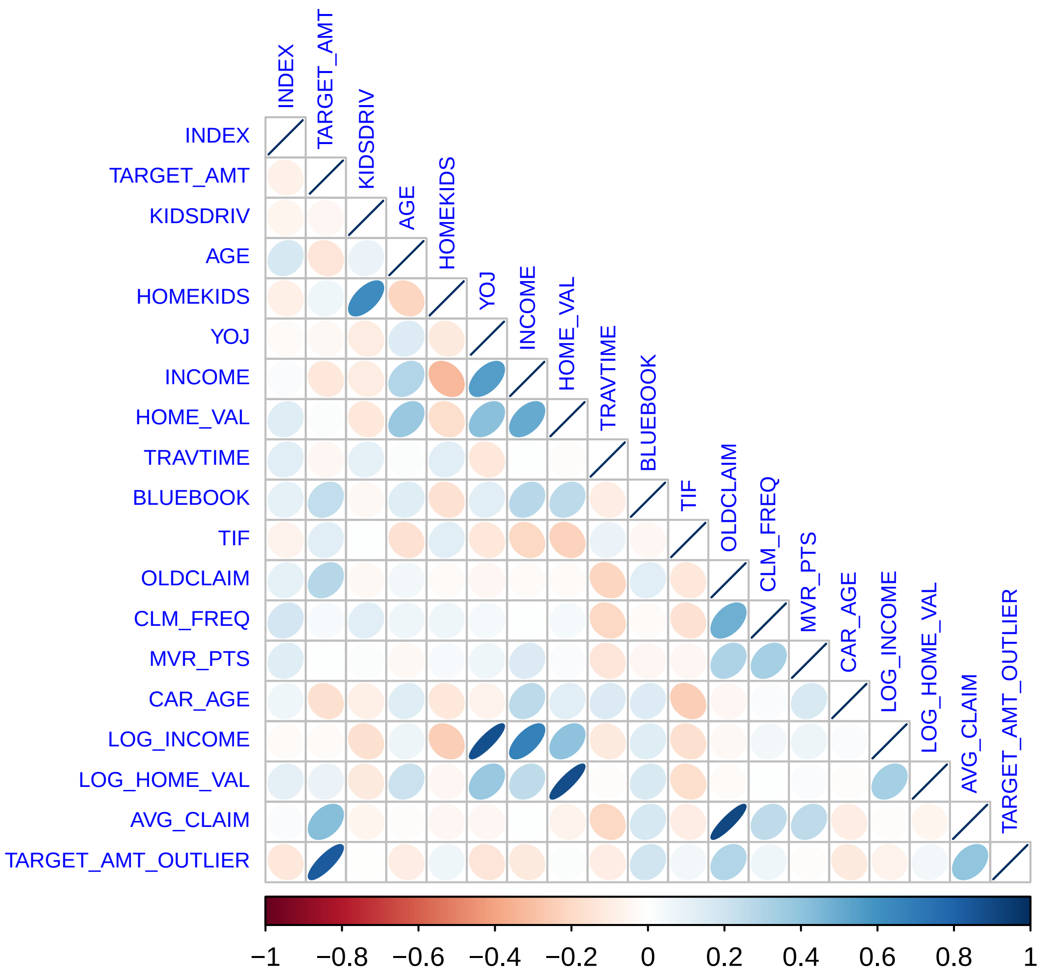



```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK, data = .)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7277  -2969  -1480    340  71998
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.197e+03  3.785e+02  11.089  < 2e-16 ***
## BLUEBOOK    9.670e-02  2.258e-02   4.283 1.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7110 on 1420 degrees of freedom
## Multiple R-squared:  0.01275,    Adjusted R-squared:  0.01206
```

```
## F-statistic: 18.35 on 1 and 1420 DF,  p-value: 1.966e-05
```

Rural Model

Let's filter by RURAL_DRIVER, (URBAN_DRIVER=0):




```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK + CAR_AGE + TRAVTIME, data = .)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8930  -2148   -758    796  37462
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 3169.59050 1922.29728   1.649   0.1030
## BLUEBOOK       0.22747    0.08761   2.596   0.0112 *
## CAR_AGE     -257.35639  130.60535  -1.970   0.0522 .
## TRAVTIME       4.62814   35.31399   0.131   0.8961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5406 on 82 degrees of freedom
## Multiple R-squared:  0.1029, Adjusted R-squared:  0.07005
## F-statistic: 3.134 on 3 and 82 DF,  p-value: 0.02989
```

Predictions

We assume that everyone with a TARGET_FLAG = 0 has a TARGET_AMT as zero. We then refine it with the two linear models:

```
## # A tibble: 2,141 x 39
##    INDEX TARGET_FLAG TARGET~1 KIDSD~2   AGE HOMEK~3   YOJ INCOME PARENT1 HOME_~4
##    <int>       <dbl>    <dbl>   <int> <int>   <int> <int>  <dbl> <chr>     <dbl>
##  1     3           0        0       0    48       0    11  52881 No            0
##  2     9           1    6028.       1    40       1    11  50815 Yes           0
##  3    10           0        0       0    44       2    12  43486 Yes           0
##  4    18           0        0       0    35       2    NA  21204 Yes           0
##  5    21           1    5689.       0    59       0    12  87460 No            0
##  6    30          NA       NA       0    46       0    14     NA No       207519
##  7    31           1    5289.       0    60       0    12  37940 No       182739
##  8    37           1    6518.       0    54       0    12  33212 No       158432
##  9    39           0        0       2    36       2    12 130540 Yes      344195
## 10    47           0        0       0    50       0     8 167469 No            0
## # ... with 2,131 more rows, 29 more variables: MSTATUS <chr>, SEX <chr>,
## #   EDUCATION <chr>, JOB <chr>, TRAVTIME <int>, CAR_USE <chr>, BLUEBOOK <dbl>,
## #   TIF <int>, CAR_TYPE <chr>, RED_CAR <chr>, OLDCLAIM <dbl>, CLM_FREQ <int>,
## #   REVOKED <chr>, MVR_PTS <int>, CAR_AGE <int>, LOG_INCOME <dbl>,
## #   LOG_HOME_VAL <dbl>, AVG_CLAIM <dbl>, PRIOR_ACCIDENT <fct>,
## #   COLLEGE_EDUCATED <fct>, URBAN_DRIVER <fct>, YOUNG_MALE <fct>, YOUNG <fct>,
## #   RED_SPORTS_CAR <fct>, HAS_KIDS <fct>, KID_DRIVERS <fct>, ...
```

Let's predict the estimations values of the evaluation data set. Then we'll write it to csv.