DATA621-Homework4-SmoothOperators

Rob Hodde, Matt Farris, Jeffrey Burmood, Bin Lin

4/17/2017

Problem Description

The 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.

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.

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 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.

Approach Steps:

1) Build a logistic regression model based on the TARGET_FLAG response variable.
2) Generate TARGET_FLAG predictions using the logistic regression model.
3) Build a linear regression model based on the non-zero values of the TARGET_AMT response variable.
4) Generate TARGET_AMT predictions using the linear regression model based on the non-zero values of the predicted TARGET_FLAG variable.

Data Exploration

Data Exploration

```
## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

## Loading required package: bitops

## Loading required package: lattice
```

```
## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units

##
## Attaching package: 'caret'

## The following object is masked from 'package:survival':
##
##     cluster

## Loading required package: Rcpp

## mice 2.25 2015-11-09

##
## Attaching package: 'mice'

## The following object is masked from 'package:RCurl':
##
##     complete

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:stringr':
##
##     %>%

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##      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   Min.   :     0
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097
##  Median :45.00   Median :0.0000   Median :11.0   Median : 54028
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   Mean   : 61898
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   Max.   :367030
##  NA's   :6                        NA's   :454    NA's   :445
##    PARENT1             HOME_VAL        MSTATUS              SEX
##  Length:8161        Min.   :     0   Length:8161        Length:8161
##  Class :character   1st Qu.:     0   Class :character   Class :character
##  Mode  :character   Median :161160   Mode  :character   Mode  :character
##                     Mean   :154867
##                     3rd Qu.:238724
##                     Max.   :885282
##                     NA's   :464
##   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
##  Min.   : 1500   Min.   : 1.000   Length:8161        Length:8161
##  1st Qu.: 9280   1st Qu.: 1.000   Class :character   Class :character
##  Median :14440   Median : 4.000   Mode  :character   Mode  :character
##  Mean   :15710   Mean   : 5.351
##  3rd Qu.:20850   3rd Qu.: 7.000
##  Max.   :69740   Max.   :25.000
##
##     OLDCLAIM        CLM_FREQ        REVOKED             MVR_PTS
##  Min.   :    0   Min.   :0.0000   Length:8161        Min.   : 0.000
##  1st Qu.:    0   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Median :    0   Median :0.0000   Mode  :character   Median : 1.000
##  Mean   : 4037   Mean   :0.7986                      Mean   : 1.696
##  3rd Qu.: 4636   3rd Qu.:2.0000                      3rd Qu.: 3.000
##  Max.   :57037   Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE         URBANICITY          blnPARENT1      blnMSTATUS
##  Min.   : 0.000   Length:8161        Min.   :0.000   Min.   :0.0000
##  1st Qu.: 1.000   Class :character   1st Qu.:0.000   1st Qu.:0.0000
##  Median : 8.000   Mode  :character   Median :0.000   Median :1.0000
##  Mean   : 8.329                      Mean   :0.132   Mean   :0.5997
##  3rd Qu.:12.000                      3rd Qu.:0.000   3rd Qu.:1.0000
```

```
##  Max.   :28.000                     Max.   :1.000   Max.   :1.0000
##  NA's   :510
##      blnSEX         blnCAR_USE     blnNOT_RED_CAR   blnNOT_REVOKED
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.5361   Mean   :0.6288   Mean   :0.7086   Mean   :0.8775
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##  blnURBANICITY     intEDUCATION       intJOB       intCAR_TYPE
##  Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.0000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000
##  Median :1.0000   Median :3.000   Median :4.000   Median :3.000
##  Mean   :0.7955   Mean   :2.801   Mean   :4.248   Mean   :3.192
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:6.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000   Max.   :6.000
##                                   NA's   :526
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME HOME_VAL
## 1     1           0          0        0  60        0  11  67349        0
## 2     2           0          0        0  43        0  11  91449   257252
## 3     4           0          0        0  35        1  10  16039   124191
## 4     5           0          0        0  51        0  14  51497   306251
## 5     6           0          0        0  50        0  14 114986   243925
## 6     7           1       2946        0  34        1  12 125301        0
##   TRAVTIME BLUEBOOK TIF OLDCLAIM CLM_FREQ MVR_PTS CAR_AGE blnPARENT1
## 1       14    14230  11     4461        2       3      18          0
## 2       22    14940   1        0        0       0       1          0
## 3        5     4010   4    38690        2       3      10          0
## 4       32    15440   7        0        0       0       6          0
## 5       36    18000   1    19217        2       3      17          0
## 6       46    17430   1        0        0       0       7          1
##   blnMSTATUS blnSEX blnCAR_USE blnNOT_RED_CAR blnNOT_REVOKED blnURBANICITY
## 1          0      0          1              0              1             1
## 2          0      0          0              0              1             1
## 3          1      1          1              1              1             1
## 4          1      0          1              0              1             1
## 5          1      1          1              1              0             1
## 6          0      1          0              1              1             1
##   intEDUCATION intJOB intCAR_TYPE
## 1            5      5           4
## 2            2      4           4
## 3            2      3           2
## 4            1      4           4
## 5            5      8           2
## 6            3      4           1
```

```
##      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   Min.   :     0
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 27907
##  Median :45.00   Median :0.0000   Median :11.0   Median : 53564
##  Mean   :44.78   Mean   :0.7212   Mean   :10.5   Mean   : 61582
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85479
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   Max.   :367030
##     HOME_VAL         TRAVTIME         BLUEBOOK          TIF
##  Min.   :     0   Min.   :  5.00   Min.   : 1500   Min.   : 1.000
##  1st Qu.:     0   1st Qu.: 22.00   1st Qu.: 9280   1st Qu.: 1.000
##  Median :162136   Median : 33.00   Median :14440   Median : 4.000
##  Mean   :155482   Mean   : 33.49   Mean   :15710   Mean   : 5.351
##  3rd Qu.:239130   3rd Qu.: 44.00   3rd Qu.:20850   3rd Qu.: 7.000
##  Max.   :885282   Max.   :142.00   Max.   :69740   Max.   :25.000
##     OLDCLAIM        CLM_FREQ         MVR_PTS          CAR_AGE
##  Min.   :    0   Min.   :0.0000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:    0   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :    0   Median :0.0000   Median : 1.000   Median : 8.000
##  Mean   : 4037   Mean   :0.7986   Mean   : 1.696   Mean   : 8.339
##  3rd Qu.: 4636   3rd Qu.:2.0000   3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :57037   Max.   :5.0000   Max.   :13.000   Max.   :28.000
##    blnPARENT1      blnMSTATUS         blnSEX         blnCAR_USE
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.132   Mean   :0.5997   Mean   :0.5361   Mean   :0.6288
##  3rd Qu.:0.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  blnNOT_RED_CAR   blnNOT_REVOKED   blnURBANICITY     intEDUCATION
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:2.000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :3.000
##  Mean   :0.7086   Mean   :0.8775   Mean   :0.7955   Mean   :2.801
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :5.000
##      intJOB       intCAR_TYPE
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000
##  Median :4.000   Median :3.000
##  Mean   :4.375   Mean   :3.192
##  3rd Qu.:6.000   3rd Qu.:4.000
##  Max.   :8.000   Max.   :6.000
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME HOME_VAL
## 1     1           0          0        0  60        0  11  67349        0
## 2     2           0          0        0  43        0  11  91449   257252
## 3     4           0          0        0  35        1  10  16039   124191
## 4     5           0          0        0  51        0  14  51497   306251
## 5     6           0          0        0  50        0  14 114986   243925
## 6     7           1       2946        0  34        1  12 125301        0
##   TRAVTIME BLUEBOOK TIF OLDCLAIM CLM_FREQ MVR_PTS CAR_AGE blnPARENT1
## 1       14    14230  11     4461        2       3      18          0
## 2       22    14940   1        0        0       0       1          0
```

```
## 3        5     4010   4    38690        2       3      10          0
## 4       32    15440   7        0        0       0       6          0
## 5       36    18000   1    19217        2       3      17          0
## 6       46    17430   1        0        0       0       7          1
##   blnMSTATUS blnSEX blnCAR_USE blnNOT_RED_CAR blnNOT_REVOKED blnURBANICITY
## 1          0      0          1              0              1             1
## 2          0      0          0              0              1             1
## 3          1      1          1              1              1             1
## 4          1      0          1              0              1             1
## 5          1      1          1              1              0             1
## 6          0      1          0              1              1             1
##   intEDUCATION intJOB intCAR_TYPE
## 1            5      5           4
## 2            2      4           4
## 3            2      3           2
## 4            1      4           4
## 5            5      8           2
## 6            3      4           1
```

Below is a summary of each predictor variable's basic statistics, followed by boxplots which illustrate the spread and outliers for each variable.

VAR	TYPE
TARGET_FLAG	double
TARGET_AMT	double
KIDSDRIV	integer
AGE	integer
HOMEKIDS	integer
YOJ	integer
INCOME	double
HOME_VAL	double
TRAVTIME	integer
BLUEBOOK	double
TIF	integer
OLDCLAIM	double
CLM_FREQ	integer
MVR_PTS	integer
CAR_AGE	double
blnPARENT1	double
blnMSTATUS	double
blnSEX	double
blnCAR_USE	double
blnNOT_RED_CAR	double
blnNOT_REVOKED	double

VAR	TYPE
blnURBANICITY	double
intEDUCATION	double
intJOB	double
intCAR_TYPE	double

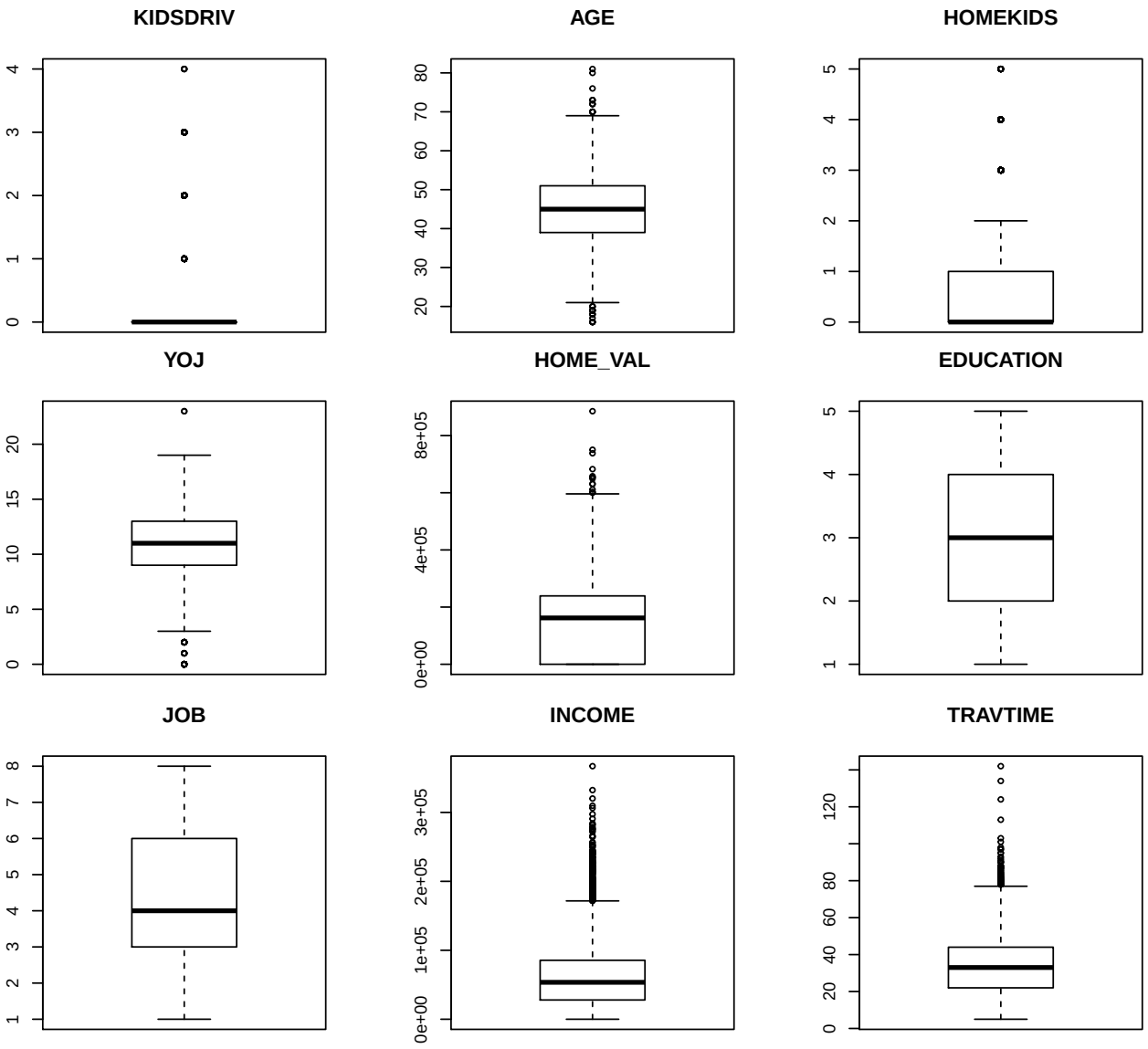
TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ
Min. :0.0000	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000	Min. : 0.0
1st Qu.:0.0000	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000	1st Qu.: 9.0
Median :0.0000	Median : 0	Median :0.0000	Median :45.00	Median :0.0000	Median :11.0
Mean :0.2638	Mean : 1504	Mean :0.1711	Mean :44.78	Mean :0.7212	Mean :10.5
3rd Qu.:1.0000	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:13.0
Max. :1.0000	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000	Max. :23.0

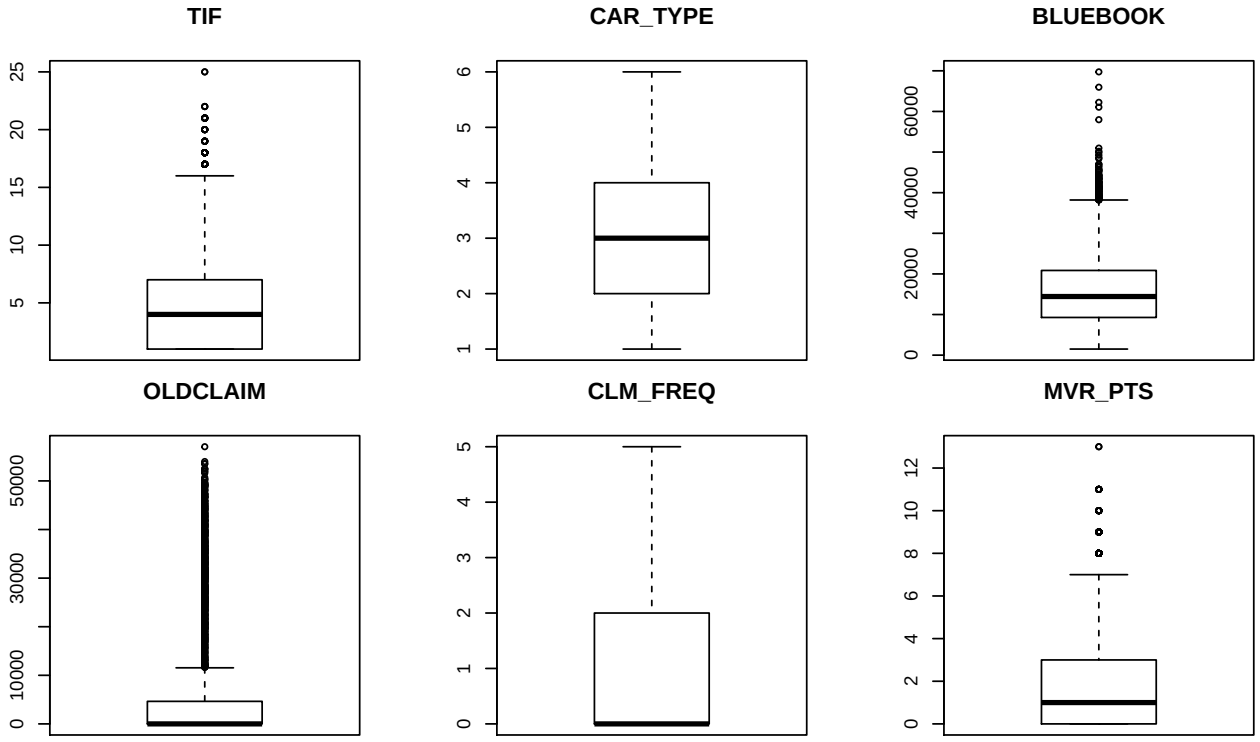
INCOME	HOME_VAL	TRAVTIME	BLUEBOOK	TIF	OLDCLAIM
Min. : 0	Min. : 0	Min. : 5.00	Min. : 1500	Min. : 1.000	Min. : 0
1st Qu.: 27907	1st Qu.: 0	1st Qu.: 22.00	1st Qu.: 9280	1st Qu.: 1.000	1st Qu.: 0
Median : 53564	Median :162136	Median : 33.00	Median :14440	Median : 4.000	Median : 0
Mean : 61582	Mean :155482	Mean : 33.49	Mean :15710	Mean : 5.351	Mean : 4037
3rd Qu.: 85479	3rd Qu.:239130	3rd Qu.: 44.00	3rd Qu.:20850	3rd Qu.: 7.000	3rd Qu.: 4636
Max. :367030	Max. :885282	Max. :142.00	Max. :69740	Max. :25.000	Max. :57037

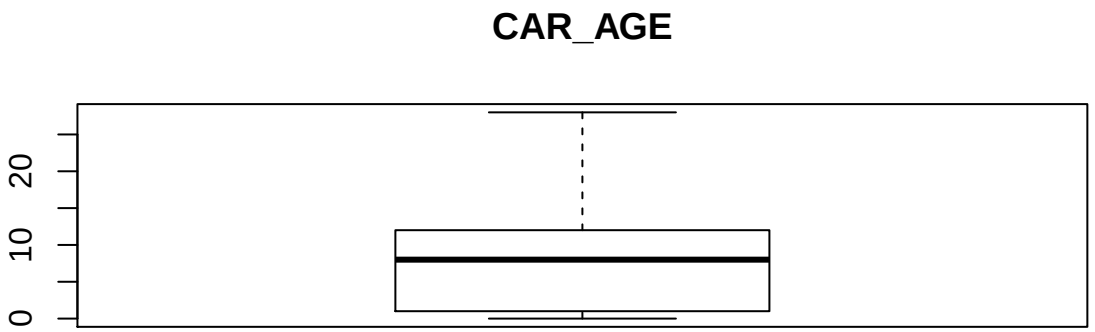
CLM_FREQ	MVR_PTS	CAR_AGE	blnPARENT1	blnMSTATUS	blnSEX
Min. :0.0000	Min. : 0.000	Min. : 0.000	Min. :0.000	Min. :0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.: 0.000	1st Qu.: 1.000	1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.0000	Median : 1.000	Median : 8.000	Median :0.000	Median :1.0000	Median :1.0000
Mean :0.7986	Mean : 1.696	Mean : 8.339	Mean :0.132	Mean :0.5997	Mean :0.5361
3rd Qu.:2.0000	3rd Qu.: 3.000	3rd Qu.:12.000	3rd Qu.:0.000	3rd Qu.:1.0000	3rd Qu.:1.0000
Max. :5.0000	Max. :13.000	Max. :28.000	Max. :1.000	Max. :1.0000	Max. :1.0000

blnCAR_USE	blnNOT_RED_CAR	blnNOT_REVOKED	blnURBANICITY	intEDUCATION	intJOB
Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :1.000	Min. :1.000
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:1.0000	1st Qu.:1.0000	1st Qu.:2.000	1st Qu.:3.000
Median :1.0000	Median :1.0000	Median :1.0000	Median :1.0000	Median :3.000	Median :4.000

blnCAR_USE	blnNOT_RED_CAR	blnNOT_REVOKED	blnURBANICITY	intEDUCATION	intJOB
Mean :0.6288	Mean :0.7086	Mean :0.8775	Mean :0.7955	Mean :2.801	Mean :4.375
3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:4.000	3rd Qu.:6.000
Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :5.000	Max. :8.000







	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME
KIDSDRIV	1.0000000	-0.0747660	0.4640152	0.0432964	-0.0455581	-0.0221674	0.0084473
AGE	-0.0747660	1.0000000	-0.4460301	0.1347209	0.1808708	0.2137132	0.0058049
HOMEKIDS	0.4640152	-0.4460301	1.0000000	0.0868267	-0.1591347	-0.1125173	-0.0072456
YOJ	0.0432964	0.1347209	0.0868267	1.0000000	0.2836622	0.2719705	-0.0171929
INCOME	-0.0455581	0.1808708	-0.1591347	0.2836622	1.0000000	0.5840194	-0.0478402
HOME_VAL	-0.0221674	0.2137132	-0.1125173	0.2719705	0.5840194	1.0000000	-0.0394062
TRAVTIME	0.0084473	0.0058049	-0.0072456	-0.0171929	-0.0478402	-0.0394062	1.0000000
BLUEBOOK	-0.0215493	0.1657139	-0.1078936	0.1418059	0.4317290	0.2573569	-0.0170013
TIF	-0.0019887	-0.0003216	0.0118133	0.0250877	0.0002621	0.0055782	-0.0116046
OLDCLAIM	0.0204027	-0.0293812	0.0299110	-0.0018460	-0.0428910	-0.0698640	-0.0192672
CLM_FREQ	0.0370629	-0.0241887	0.0293493	-0.0274784	-0.0452202	-0.0915493	0.0065602
MVR_PTS	0.0535664	-0.0722675	0.0606013	-0.0375184	-0.0584554	-0.0796242	0.0105985
CAR_AGE	-0.0517842	0.1798145	-0.1533148	0.0601073	0.4132529	0.2251718	-0.0402656

	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME
blnPARENT1	0.1966038	-0.3150435	0.4492740	-0.0499350	-0.0720065	-0.2593940	-0.0237406
blnMSTATUS	0.0424609	0.0912595	0.0435259	0.1447036	-0.0328041	0.4566589	0.0102483
blnSEX	0.0459344	-0.0661930	0.1115114	-0.0803406	-0.1119631	-0.0736697	0.0046177
blnCAR_USE	-0.0014216	0.0328987	0.0044583	-0.0215889	-0.0847237	-0.0291608	-0.0248054
blnNOT_RED_CAR	0.0436382	-0.0186567	0.0681480	-0.0466129	-0.0618827	-0.0138569	-0.0039658
blnNOT_REVOKED	-0.0430620	0.0384171	-0.0451156	0.0040409	0.0194742	0.0493280	0.0121153
blnURBANICITY	-0.0371236	0.0511078	-0.0634829	0.0821629	0.2081851	0.1231965	-0.1660047
intEDUCATION	-0.0714891	0.2448180	-0.2036956	0.0839070	0.6055095	0.3508302	-0.0572046
intJOB	-0.0706557	0.2505999	-0.2223849	0.3298813	0.6848735	0.4449778	-0.0842505
intCAR_TYPE	-0.0317930	0.0452547	-0.1052492	0.1018265	0.2835755	0.1626513	-0.0113547

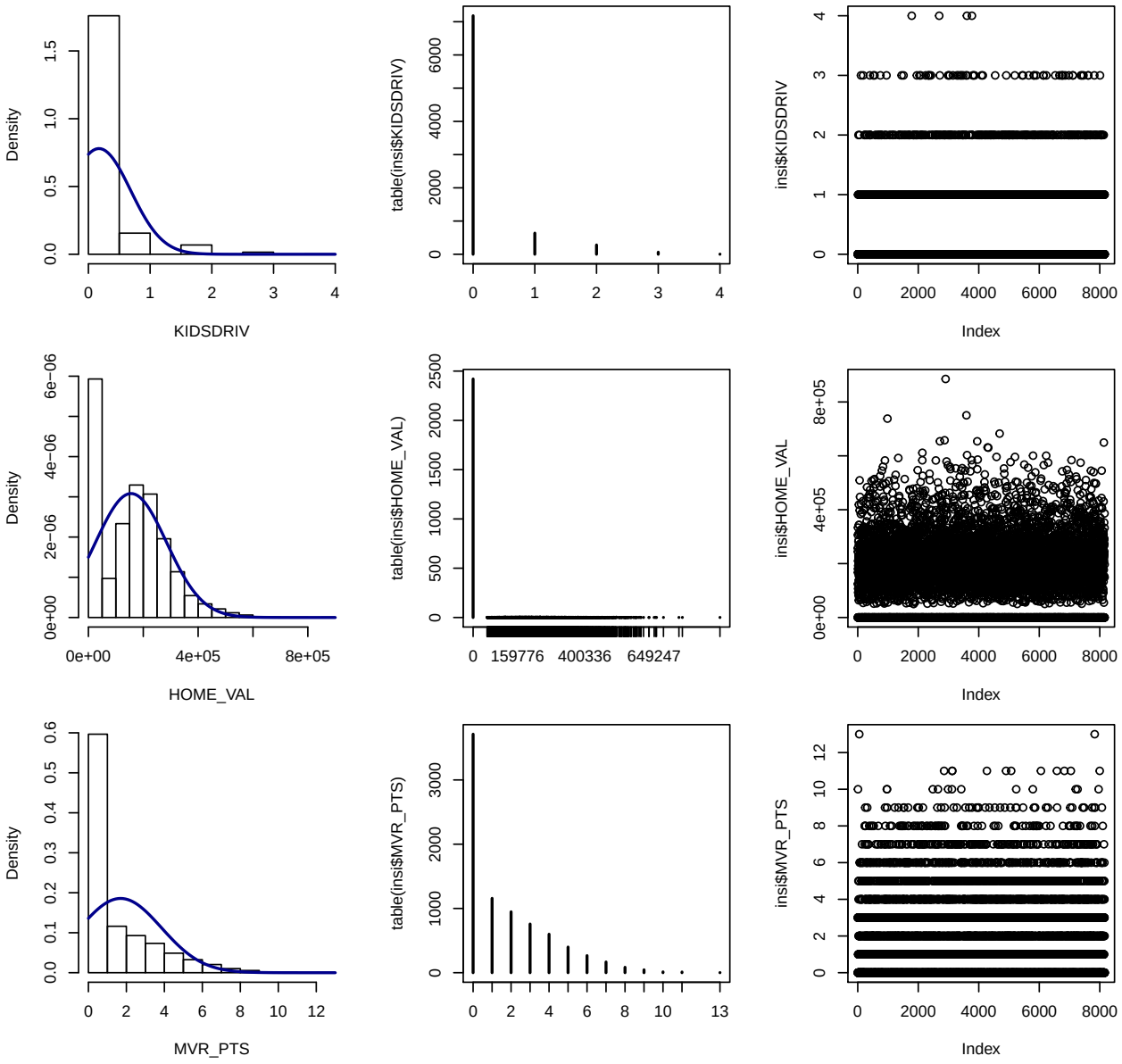
	TIF	OLDCLAIM	CLM_FREQ	MVR_PTS	CAR_AGE	blnPARENT1	blnMSTA
KIDSDRIV	-0.0019887	0.0204027	0.0370629	0.0535664	-0.0517842	0.1966038	0.042
AGE	-0.0003216	-0.0293812	-0.0241887	-0.0722675	0.1798145	-0.3150435	0.091
HOMEKIDS	0.0118133	0.0299110	0.0293493	0.0606013	-0.1533148	0.4492740	0.043
YOJ	0.0250877	-0.0018460	-0.0274784	-0.0375184	0.0601073	-0.0499350	0.144
INCOME	0.0002621	-0.0428910	-0.0452202	-0.0584554	0.4132529	-0.0720065	-0.032
HOME_VAL	0.0055782	-0.0698640	-0.0915493	-0.0796242	0.2251718	-0.2593940	0.456
TRAVTIME	-0.0116046	-0.0192672	0.0065602	0.0105985	-0.0402656	-0.0237406	0.010
BLUEBOOK	-0.0054246	-0.0295176	-0.0363415	-0.0391308	0.1863770	-0.0504582	-0.007
TIF	1.0000000	-0.0219582	-0.0230230	-0.0410457	0.0097276	-0.0019519	-0.000
OLDCLAIM	-0.0219582	1.0000000	0.4951308	0.2644850	-0.0130352	0.0346893	-0.045
CLM_FREQ	-0.0230230	0.4951308	1.0000000	0.3966384	-0.0080839	0.0487424	-0.069
MVR_PTS	-0.0410457	0.2644850	0.3966384	1.0000000	-0.0150483	0.0684526	-0.047
CAR_AGE	0.0097276	-0.0130352	-0.0080839	-0.0150483	1.0000000	-0.0614436	-0.030
blnPARENT1	-0.0019519	0.0346893	0.0487424	0.0684526	-0.0614436	1.0000000	-0.477
blnMSTATUS	-0.0007411	-0.0459198	-0.0693289	-0.0479670	-0.0309293	-0.4772281	1.000
blnSEX	-0.0061012	0.0000909	-0.0122335	0.0073444	-0.0190691	0.0737837	0.004
blnCAR_USE	-0.0001161	-0.0357676	-0.0814907	-0.0680838	0.0663192	-0.0061940	0.020
blnNOT_RED_CAR	0.0008717	-0.0138214	-0.0260815	-0.0060406	-0.0194762	0.0420856	0.019
blnNOT_REVOKED	0.0318415	-0.4181035	-0.0530499	-0.0531731	0.0072316	-0.0497187	0.043
blnURBANICITY	0.0071310	0.1510826	0.2391246	0.1502433	0.1662379	-0.0222096	-0.002
intEDUCATION	0.0019466	-0.0234187	-0.0126025	-0.0338609	0.6903594	-0.0815011	-0.039
intJOB	0.0073937	-0.0272436	-0.0194324	-0.0349784	0.4898766	-0.0920789	-0.027
intCAR_TYPE	0.0010737	-0.0260373	-0.0209609	-0.0263118	0.1075188	-0.0590901	-0.016

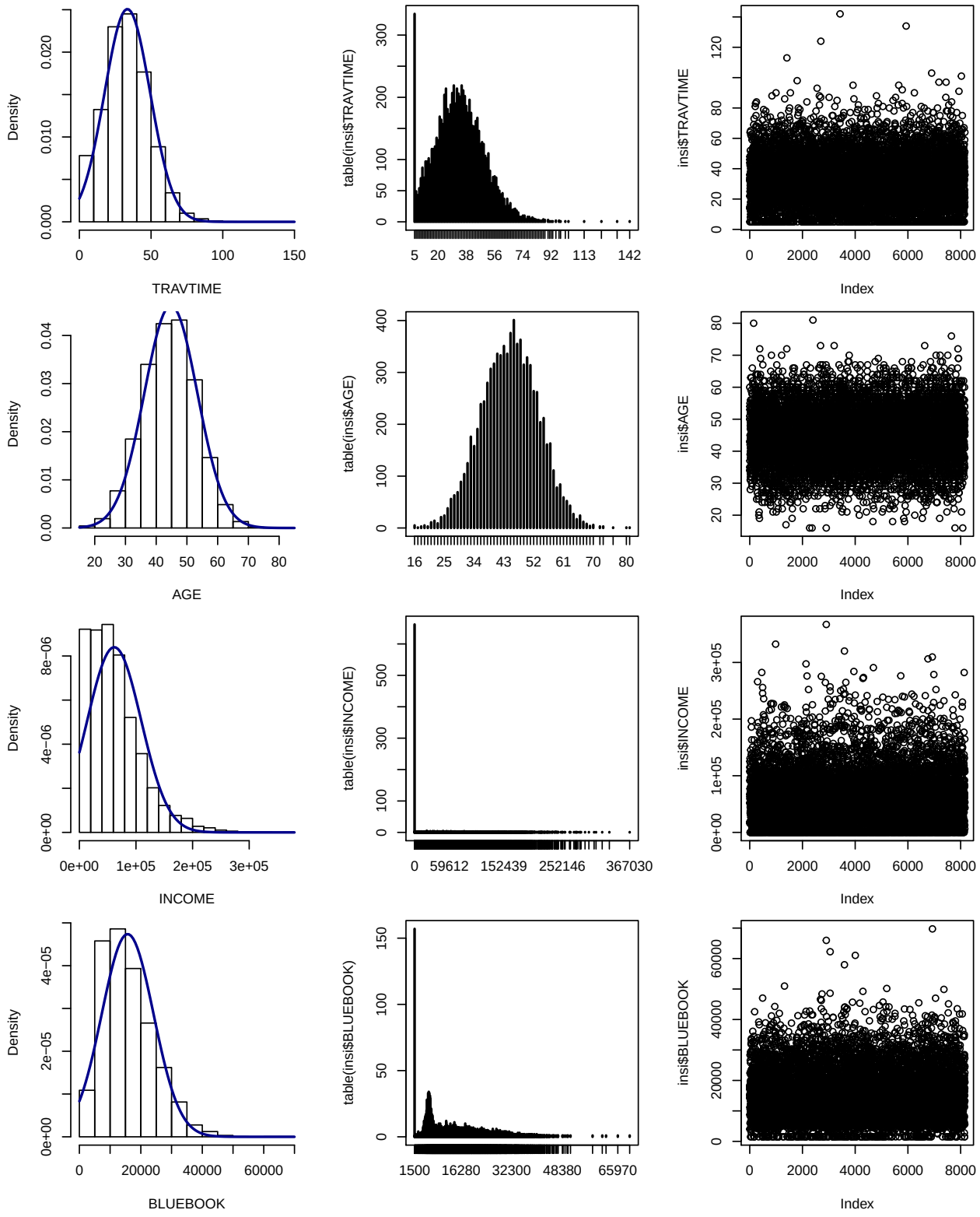
Here are the results from an analysis of the predictor variable correlations:

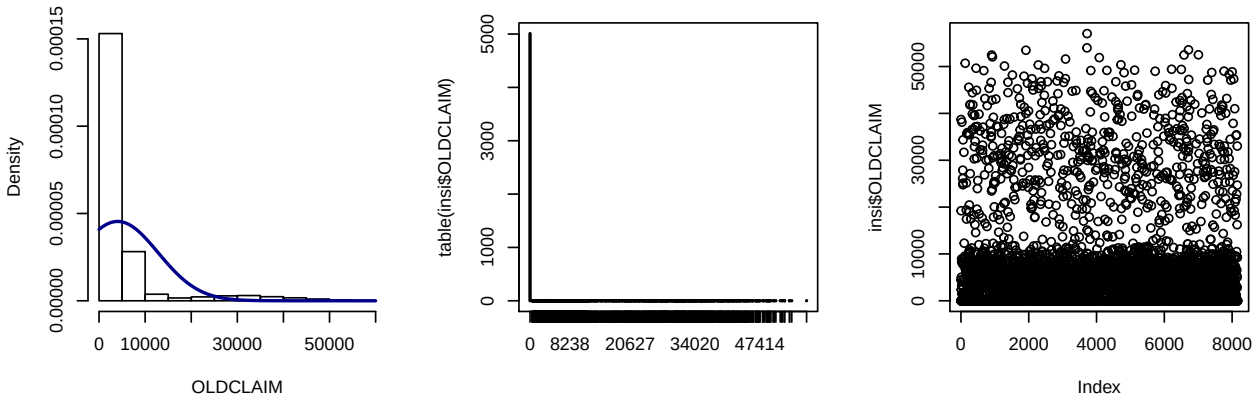
The are no strong correlations (>70%) between predictor variables, not enough to allow consideration of removing a variable from the model based on a high correlation with another variable. There is some moderate correlation (30-50%) between some variable highlighting obvious relationships such as HOMEKIDS-KIDSDRIV, HOME_VAL-INCOME, EDUCATION-INCOME, JOB-INCOME, CAR_TYPE-BLUEBOOK, CLM_FREQ-OLDCLAIM, and MVR_PTS-CLM_FREQ.

Based on an analysis of the box plots, the following variables have some outliers that may, or may not, exert influence on the regression results: - KIDSDRIV, HOME_VAL, TRAVTIME, MVR_PTS, AGE, INCOME, BLUEBOOK, OLDCLAIM

We'll next look at these variables more closely, starting with their histograms and frequency counts to better understand the nature of their distribution.




Density
TRAVTIME
table(insi$TRAVTIME)
insi$TRAVTIME
Index
Density
AGE
table(insi$AGE)
insi$AGE
Index
Density
INCOME
table(insi$INCOME)
insi$INCOME
Index
Density
BLUEBOOK
table(insi$BLUEBOOK)
insi$BLUEBOOK
Index



The analysis of the distributions for these variables show varying degrees of skewness, except for the AGE variable, which shows a fairly normal distribution.

For the logistic regression analysis, we would like to remove as much of the skewness as possible from the candidate predictor variables. A transformation analysis was performed and a log transformation of most of the skewed variables result in a near-normal distribution consequently, for the logistic regression model, we will use the log value of the following variables for the modeling: OLDCLAIM, BLUEBOOK, TRAVTIME, HOME_VAL.

The last step in the data exploration is to examine the correlation tables, to see if there is any potential correlations that will impact our models. The below table shows the a condense correlation table showing only the highest positive and negative correlations (in this case above 0.5 and below -.5)

As you can see, certain variables are correlated, and they are values that we would expect to be correlated. We expect there to be a correlation between income and jobs/home value and educations, as these are well established indicators of success. Higher educated people are more likely to secure higher paying jobs, which we can see in the higher correlations for these values. Furthermore, car type is going to be a clear indicator of the value of the car. Sports cars and suvs cost more than compacts. An interestiong but also intuitive correlations exists bewteen gender and car type and car color. The car type is very interesting, as we almost arbitraily set the type to a numeric value 1 - 6 according to this list “Sports Car”,“SUV”,“Pickup”,“Minivan”,“Van”,“Panel Truck” respectively. As the correlation is negative, it shows that as the value goes “up/higher” we see less women in these types of cars. Though this is a binary correlation, and the weight of such can be consider not very conclusive, it is interesting to note. The most likely reason for this offset correlation is because panel truck is listed highest, which we can assume would be a predominately male oriented mode of transportation.

Data Preparation

Data Preparation

Here we do the log transformation of the variables identifed earlier with high skewness. One of the variables, OLDCLAIM, has such a distorted distribution that the log transformation will not be sufficient to make the variable a viable candidate for our model. For the OLDCLAIM variable, the difference between the median and the mean is so large we will not attempt to use the OLDCLAIM variable.

Next a brief review of the linear models of all the variables

```
##              var        p_val        r2_val
## 1        KIDSDRIV 9.992952e-01 -0.0004648997
## 2             AGE 1.877176e-01  0.0003421432
```

```
## 3        HOMEKIDS 9.826480e-01 -0.0004646800
## 4             YOJ 9.794585e-02  0.0008083879
## 5          INCOME 3.088518e-02  0.0017003885
## 6        TRAVTIME 8.822314e-01 -0.0004546904
## 7        BLUEBOOK 1.044112e-08  0.0146591269
## 8             TIF 7.803982e-01 -0.0004287390
## 9        OLDCLAIM 7.930893e-01 -0.0004328923
## 10       CLM_FREQ 9.275324e-01 -0.0004610518
## 11        MVR_PTS 6.475654e-02  0.0011208009
## 12        CAR_AGE 6.916553e-01 -0.0003917233
## 13     blnPARENT1 2.664010e-01  0.0001095732
## 14     blnMSTATUS 1.052026e-01  0.0007554980
## 15         blnSEX 1.673273e-02  0.0021945924
## 16     blnCAR_USE 2.072129e-02  0.0020209949
## 17 blnNOT_RED_CAR 2.048805e-01  0.0002825468
## 18 blnNOT_REVOKED 8.637471e-02  0.0009022378
## 19  blnURBANICITY 8.228342e-01 -0.0004415778
## 20   intEDUCATION 1.230697e-01  0.0006407052
## 21         intJOB 8.671375e-02  0.0008992983
## 22    intCAR_TYPE 3.648708e-04  0.0054278941
```

Build Models

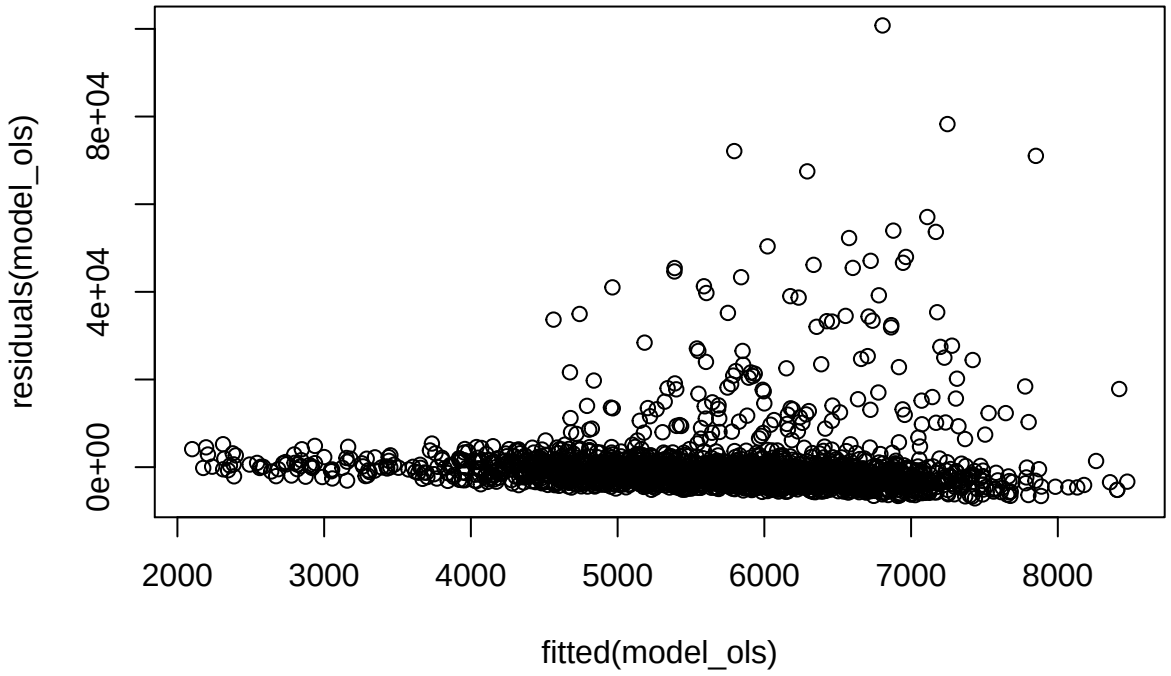
Build Models

One method of developing multiple regression models is to take a stepwise approach. To accomplish this, we combine our knowledge from the data exploration above with logistic regression. Univariate Logistic Regression is a useful method to understand how each predictor variable interacts individually with the target (response) variable. Looking at various statistics, we determine which variable may impact our target the most.

Linear Regression Models

```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + OLDCLAIM +
##     YOJ + intEDUCATION + intJOB + CAR_AGE, data = lin_model)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7119  -3088  -1556    299 100780
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.945e+03  2.662e+03  -2.984  0.00288 **
## INCOME       -7.994e-04  6.007e-03  -0.133  0.89415
## TRAVTIME     -4.219e+01  2.956e+02  -0.143  0.88654
## BLUEBOOK      1.431e+03  2.718e+02   5.266 1.54e-07 ***
## OLDCLAIM     -9.606e-04  1.649e-02  -0.058  0.95354
## YOJ           2.815e+01  4.332e+01   0.650  0.51587
## intEDUCATION  3.726e+02  2.441e+02   1.526  0.12707
## intJOB       -4.626e+01  1.508e+02  -0.307  0.75896
```

```
## CAR_AGE      -8.520e+01  4.192e+01  -2.033  0.04221 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7690 on 2144 degrees of freedom
## Multiple R-squared:  0.01733,    Adjusted R-squared:  0.01366
## F-statistic: 4.726 on 8 and 2144 DF,  p-value: 9.181e-06
```

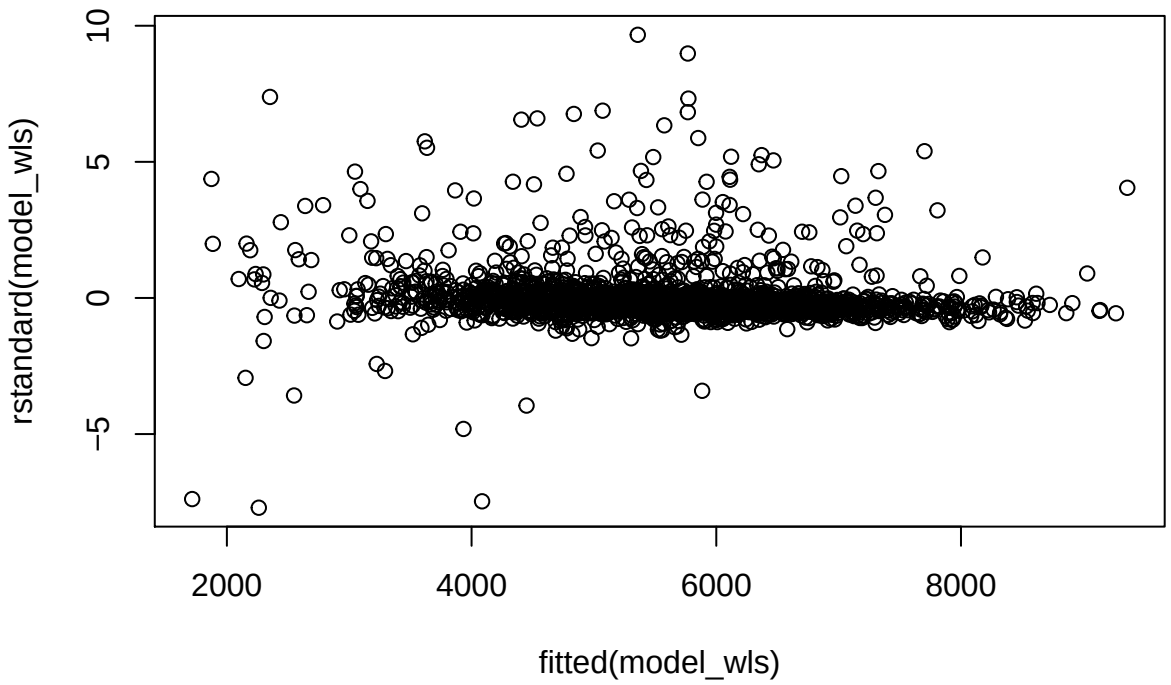



```
## [1] 2153
```

```
## [1] 2153
```

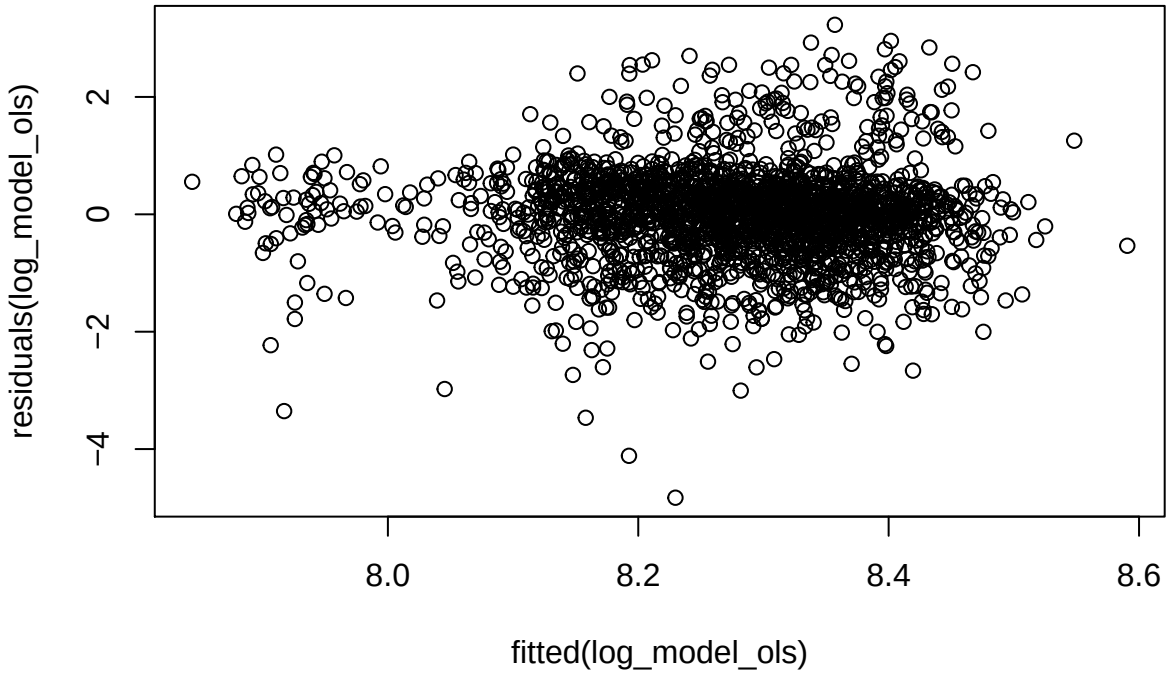
```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + OLDCLAIM +
##     YOJ + intEDUCATION + intJOB + CAR_AGE, data = lin_model,
##     weights = wts)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2291  -0.8291  -0.3839   0.1418  20.6492
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.102e+03  6.801e+02 -11.913  < 2e-16 ***
## INCOME        7.861e-04  2.282e-03   0.345 0.730487
## TRAVTIME      2.919e+02  5.526e+01   5.283 1.40e-07 ***
## BLUEBOOK      1.330e+03  8.877e+01  14.977  < 2e-16 ***
## OLDCLAIM      3.562e-02  7.771e-03   4.584 4.83e-06 ***
## YOJ           7.265e+01  1.667e+01   4.358 1.37e-05 ***
## intEDUCATION  2.035e+02  1.178e+02   1.728 0.084125 .
```

```
## intJOB       -4.979e+02  5.887e+01  -8.457  < 2e-16 ***
## CAR_AGE       8.972e+01  2.385e+01   3.761 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.137 on 2144 degrees of freedom
## Multiple R-squared:  0.1818, Adjusted R-squared:  0.1787
## F-statistic: 59.54 on 8 and 2144 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + OLDCLAIM +
##     YOJ + intEDUCATION + intJOB + CAR_AGE, data = log_lin_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8285 -0.3970  0.0426  0.3900  3.2290
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.1137965  0.5587856   9.152  < 2e-16 ***
## INCOME        0.0129460  0.0087022   1.488    0.137
## TRAVTIME     -0.0181675  0.0310089  -0.586    0.558
## BLUEBOOK      1.4257687  0.2539418   5.615 2.23e-08 ***
## OLDCLAIM      0.0006858  0.0039651   0.173    0.863
## YOJ          -0.0073611  0.0059939  -1.228    0.220
## intEDUCATION  0.0184502  0.0247647   0.745    0.456
## intJOB       -0.0124691  0.0160164  -0.779    0.436
## CAR_AGE      -0.0016229  0.0043966  -0.369    0.712
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.8064 on 2144 degrees of freedom
## Multiple R-squared:  0.01872,    Adjusted R-squared:  0.01505
## F-statistic: 5.112 on 8 and 2144 DF,  p-value: 2.515e-06
```

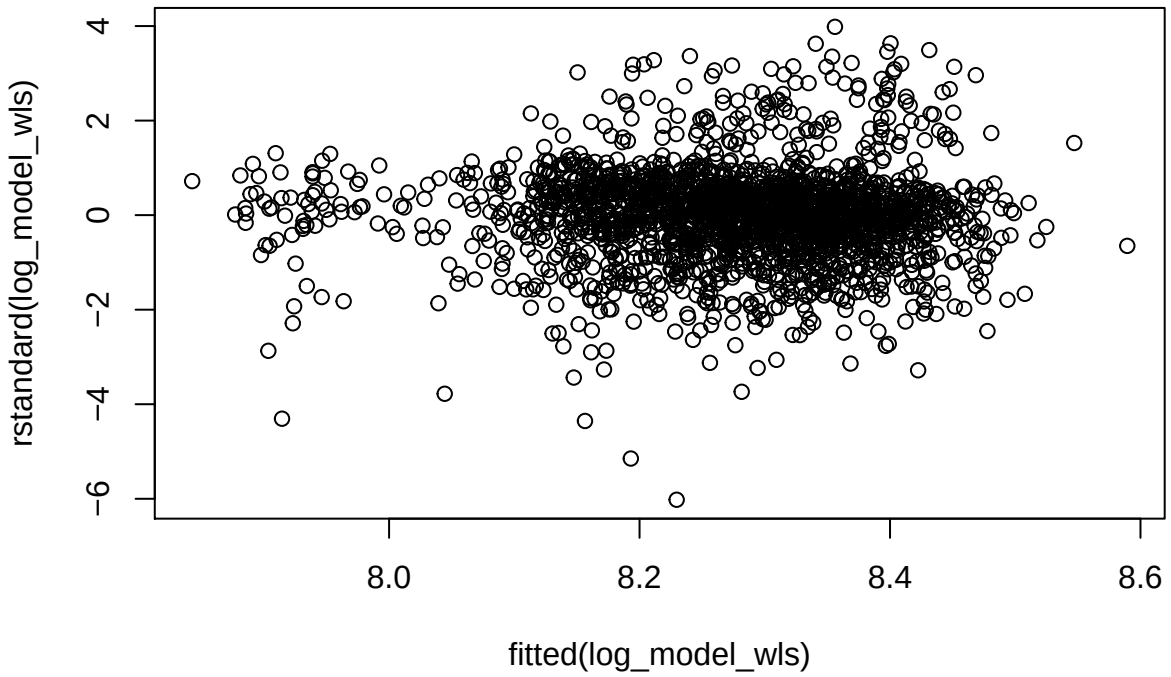



```
## [1] 2153
```

```
## [1] 2153
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + OLDCLAIM +
##     YOJ + intEDUCATION + intJOB + CAR_AGE, data = log_lin_model,
##     weights = wts)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4504 -0.6916  0.0737  0.6849  5.5913
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0988633  0.5532047   9.217  < 2e-16 ***
## INCOME        0.0128545  0.0086845   1.480    0.139
## TRAVTIME     -0.0187390  0.0309922  -0.605    0.545
## BLUEBOOK      1.4340199  0.2515170   5.701 1.35e-08 ***
## OLDCLAIM      0.0006666  0.0039625   0.168    0.866
## YOJ          -0.0073018  0.0059927  -1.218    0.223
## intEDUCATION  0.0172643  0.0247613   0.697    0.486
## intJOB       -0.0123019  0.0160140  -0.768    0.442
## CAR_AGE      -0.0014645  0.0043999  -0.333    0.739
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.405 on 2144 degrees of freedom
## Multiple R-squared:  0.01914,    Adjusted R-squared:  0.01548
## F-statistic: 5.231 on 8 and 2144 DF,  p-value: 1.681e-06
```




```
## Start:  AIC=3278.69
## TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + OLDCLAIM + YOJ +
##     intEDUCATION + intJOB + CAR_AGE
##
##                Df Sum of Sq     RSS    AIC
## - INCOME        1      0.54  9790.4 3276.8
## <none>                       9789.9 3278.7
## - intEDUCATION  1     13.64  9803.5 3279.7
## - CAR_AGE       1     64.59  9854.5 3290.8
## - YOJ           1     86.74  9876.6 3295.7
## - OLDCLAIM      1     95.94  9885.8 3297.7
## - TRAVTIME      1    127.44  9917.3 3304.5
## - intJOB        1    326.58 10116.4 3347.3
## - BLUEBOOK      1   1024.26 10814.1 3490.9
##
## Step:  AIC=40904.81
## TARGET_AMT ~ TRAVTIME + BLUEBOOK + OLDCLAIM + YOJ + intEDUCATION +
##     intJOB + CAR_AGE


##
## Call:
## lm(formula = TARGET_AMT ~ TRAVTIME + BLUEBOOK + OLDCLAIM + YOJ +
##     intEDUCATION + intJOB + CAR_AGE, data = lin_model, weights = wts)
##
## Weighted Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -12263  -5380  -2707    510 174478
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.772e+03  2.527e+03  -3.075  0.00213 **
## TRAVTIME     -4.956e+01  2.937e+02  -0.169  0.86602
## BLUEBOOK      1.418e+03  2.626e+02   5.401 7.35e-08 ***
## OLDCLAIM     -1.086e-03  1.638e-02  -0.066  0.94713
## YOJ           2.716e+01  4.229e+01   0.642  0.52086
## intEDUCATION  3.559e+02  2.341e+02   1.520  0.12860
## intJOB       -5.326e+01  1.360e+02  -0.392  0.69543
## CAR_AGE      -8.378e+01  4.171e+01  -2.009  0.04470 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13330 on 2145 degrees of freedom
## Multiple R-squared:  0.01755,    Adjusted R-squared:  0.01434
## F-statistic: 5.474 on 7 and 2145 DF,  p-value: 2.997e-06
```

Logistic Regression Model

In this model-and in all the models- we set aside 20% of the training data and use 80% to train the model we then use the model to predict the outcome of the remaining 20% of the data.

In this scenario we attempt to create the simplest model possible by using only one variable - the one that provides the highest overall AUC (performance) by itself. We calculate AUC for each variable separately and then select the highest result.

var	p_val	aic	auc
KIDSDRIV	0.0000000	7496.182	0.5326796
HOMEKIDS	0.0000000	7480.467	0.5643226
AGE	0.0000000	7489.471	0.5677821
blnPARENT1	0.0000000	7415.589	0.5618229
BLUEBOOK	0.0000000	7478.621	0.5664127
INCOME	0.0000000	7417.582	0.5924626
CAR_AGE	0.0000000	7491.509	0.5637299
intEDUCATION	0.0000000	7446.171	0.5818674
intJOB	0.0000000	7427.756	0.5851310
blnMSTATUS	0.0000000	7441.521	0.5704225
intCAR_TYPE	0.0000000	7520.419	0.5550436
CLM_FREQ	0.0000000	7272.338	0.6336553
TIF	0.0000000	7510.860	0.5383267
MVR_PTS	0.0000000	7272.810	0.6231339
blnSEX	0.0505324	7559.781	0.5050840
blnCAR_USE	0.0000000	7440.294	0.5881564

var	p_val	aic	auc
blnNOT_RED_CAR	0.5948267	7563.328	0.5044923
blnNOT_REVOKED	0.0000000	7416.317	0.5487418
blnURBANICITY	0.0000000	7163.720	0.6049599
TRAVTIME	0.0000004	7537.158	0.5239225
HOME_VAL	0.0000000	7429.881	0.5751275

We will derive our logistic regression model by stepping backward from using all candidate variables and arriving at the variable set that maximizes the AUC value.

MODEL 1 - Backward regression starting with all variables

```
## Start:  AIC=5998.19
## TARGET_FLAG ~ (TARGET_AMT + KIDSDRIV + AGE + HOMEKIDS + YOJ +
##     INCOME + HOME_VAL + TRAVTIME + BLUEBOOK + TIF + OLDCLAIM +
##     CLM_FREQ + MVR_PTS + CAR_AGE + blnPARENT1 + blnMSTATUS +
##     blnSEX + blnCAR_USE + blnNOT_RED_CAR + blnNOT_REVOKED + blnURBANICITY +
##     intEDUCATION + intJOB + intCAR_TYPE) - TARGET_AMT - OLDCLAIM
##
##                  Df Deviance    AIC
## - AGE             1   5952.3 5996.3
## - blnSEX          1   5952.4 5996.4
## - blnNOT_RED_CAR  1   5952.5 5996.5
## - YOJ             1   5953.1 5997.1
## - CAR_AGE         1   5953.1 5997.1
## - HOMEKIDS        1   5953.7 5997.7
## <none>                5952.2 5998.2
## - HOME_VAL        1   5955.0 5999.0
## - intJOB          1   5956.0 6000.0
## - intEDUCATION    1   5958.6 6002.6
## - blnPARENT1      1   5959.6 6003.6
## - INCOME          1   5961.6 6005.6
## - intCAR_TYPE     1   5967.7 6011.7
## - BLUEBOOK        1   5970.7 6014.7
## - CLM_FREQ        1   5983.2 6027.2
## - KIDSDRIV        1   5984.0 6028.0
## - TIF             1   6004.7 6048.7
## - blnMSTATUS      1   6009.3 6053.3
## - MVR_PTS         1   6011.2 6055.2
## - TRAVTIME        1   6016.0 6060.0
## - blnNOT_REVOKED  1   6025.2 6069.2
## - blnCAR_USE      1   6125.6 6169.6
## - blnURBANICITY   1   6442.2 6486.2
##
## Step:  AIC=5996.26
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + CAR_AGE +
##     blnPARENT1 + blnMSTATUS + blnSEX + blnCAR_USE + blnNOT_RED_CAR +
##     blnNOT_REVOKED + blnURBANICITY + intEDUCATION + intJOB +
```

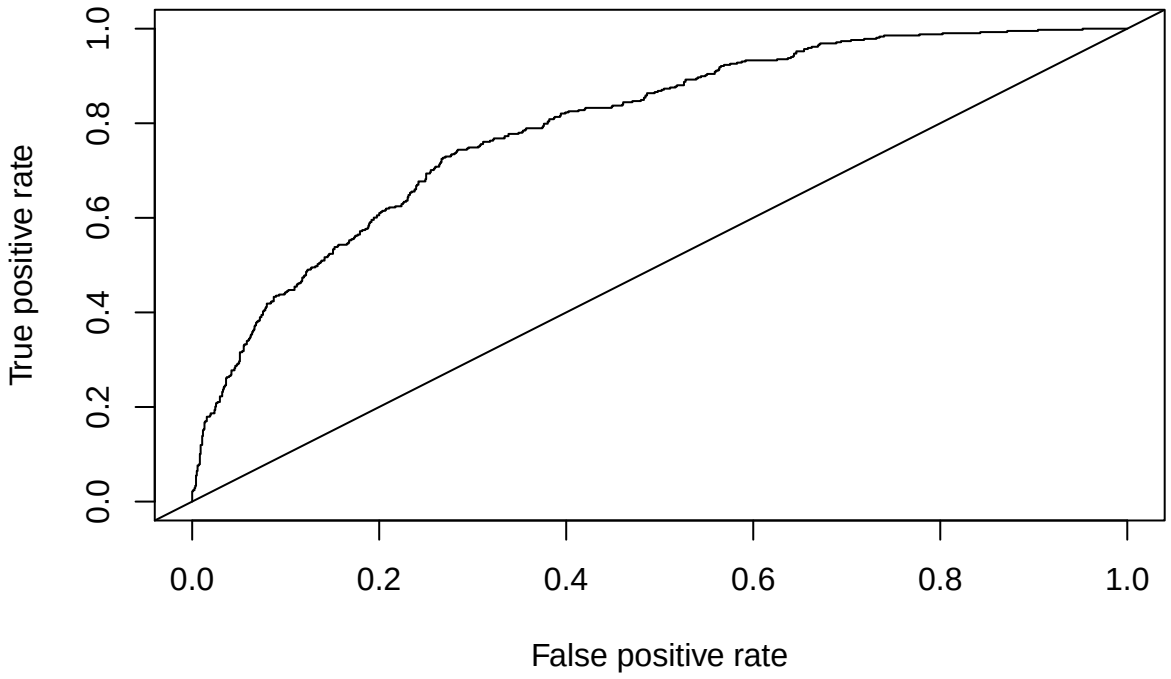
```
##     intCAR_TYPE
##
##                  Df Deviance    AIC
## - blnSEX          1   5952.5 5994.5
## - blnNOT_RED_CAR  1   5952.5 5994.5
## - CAR_AGE         1   5953.2 5995.2
## - YOJ             1   5953.2 5995.2
## <none>                5952.3 5996.3
## - HOMEKIDS        1   5954.3 5996.3
## - HOME_VAL        1   5955.1 5997.1
## - intJOB          1   5956.1 5998.1
## - intEDUCATION    1   5958.9 6000.9
## - blnPARENT1      1   5959.9 6001.9
## - INCOME          1   5961.6 6003.6
## - intCAR_TYPE     1   5967.7 6009.7
## - BLUEBOOK        1   5971.4 6013.4
## - CLM_FREQ        1   5983.3 6025.3
## - KIDSDRIV        1   5984.6 6026.6
## - TIF             1   6004.7 6046.7
## - blnMSTATUS      1   6009.5 6051.5
## - MVR_PTS         1   6011.6 6053.6
## - TRAVTIME        1   6016.0 6058.0
## - blnNOT_REVOKED  1   6025.3 6067.3
## - blnCAR_USE      1   6125.9 6167.9
## - blnURBANICITY   1   6442.9 6484.9
##
## Step:  AIC=5994.52
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + CAR_AGE +
##     blnPARENT1 + blnMSTATUS + blnCAR_USE + blnNOT_RED_CAR + blnNOT_REVOKED +
##     blnURBANICITY + intEDUCATION + intJOB + intCAR_TYPE
##
##                  Df Deviance    AIC
## - blnNOT_RED_CAR  1   5952.6 5992.6
## - CAR_AGE         1   5953.4 5993.4
## - YOJ             1   5953.5 5993.5
## <none>                5952.5 5994.5
## - HOMEKIDS        1   5954.6 5994.6
## - HOME_VAL        1   5955.5 5995.5
## - intJOB          1   5956.6 5996.6
## - intEDUCATION    1   5959.0 5999.0
## - blnPARENT1      1   5960.2 6000.2
## - INCOME          1   5961.7 6001.7
## - BLUEBOOK        1   5972.1 6012.1
## - intCAR_TYPE     1   5978.9 6018.9
## - CLM_FREQ        1   5983.5 6023.5
## - KIDSDRIV        1   5984.8 6024.8
## - TIF             1   6004.9 6044.9
## - blnMSTATUS      1   6009.7 6049.7
## - MVR_PTS         1   6011.8 6051.8
## - TRAVTIME        1   6016.3 6056.3
## - blnNOT_REVOKED  1   6025.4 6065.4
## - blnCAR_USE      1   6126.6 6166.6
## - blnURBANICITY   1   6443.0 6483.0
```

```
##
## Step:  AIC=5992.61
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + CAR_AGE +
##     blnPARENT1 + blnMSTATUS + blnCAR_USE + blnNOT_REVOKED + blnURBANICITY +
##     intEDUCATION + intJOB + intCAR_TYPE
##
##                  Df Deviance    AIC
## - CAR_AGE         1   5953.5 5991.5
## - YOJ             1   5953.6 5991.6
## <none>                5952.6 5992.6
## - HOMEKIDS        1   5954.7 5992.7
## - HOME_VAL        1   5955.5 5993.5
## - intJOB          1   5956.6 5994.6
## - intEDUCATION    1   5959.1 5997.1
## - blnPARENT1      1   5960.3 5998.3
## - INCOME          1   5961.8 5999.8
## - BLUEBOOK        1   5973.7 6011.7
## - intCAR_TYPE     1   5983.7 6021.7
## - CLM_FREQ        1   5983.7 6021.7
## - KIDSDRIV        1   5984.8 6022.8
## - TIF             1   6005.0 6043.0
## - blnMSTATUS      1   6009.9 6047.9
## - MVR_PTS         1   6011.9 6049.9
## - TRAVTIME        1   6016.4 6054.4
## - blnNOT_REVOKED  1   6025.5 6063.5
## - blnCAR_USE      1   6127.6 6165.6
## - blnURBANICITY   1   6443.0 6481.0
##
## Step:  AIC=5991.48
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + blnPARENT1 +
##     blnMSTATUS + blnCAR_USE + blnNOT_REVOKED + blnURBANICITY +
##     intEDUCATION + intJOB + intCAR_TYPE
##
##                  Df Deviance    AIC
## - YOJ             1   5954.4 5990.4
## <none>                5953.5 5991.5
## - HOMEKIDS        1   5955.6 5991.6
## - HOME_VAL        1   5956.3 5992.3
## - intJOB          1   5957.6 5993.6
## - blnPARENT1      1   5961.1 5997.1
## - INCOME          1   5962.8 5998.8
## - intEDUCATION    1   5967.1 6003.1
## - BLUEBOOK        1   5974.5 6010.5
## - CLM_FREQ        1   5984.4 6020.4
## - intCAR_TYPE     1   5984.6 6020.6
## - KIDSDRIV        1   5985.6 6021.6
## - TIF             1   6006.1 6042.1
## - blnMSTATUS      1   6010.9 6046.9
## - MVR_PTS         1   6012.8 6048.8
## - TRAVTIME        1   6017.1 6053.1
## - blnNOT_REVOKED  1   6026.6 6062.6
## - blnCAR_USE      1   6129.9 6165.9
```

```
## - blnURBANICITY   1   6444.0 6480.0
##
## Step:  AIC=5990.42
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + HOME_VAL + TRAVTIME +
##     BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + blnPARENT1 + blnMSTATUS +
##     blnCAR_USE + blnNOT_REVOKED + blnURBANICITY + intEDUCATION +
##     intJOB + intCAR_TYPE
##
##                  Df Deviance    AIC
## - HOMEKIDS        1   5956.2 5990.2
## <none>                5954.4 5990.4
## - HOME_VAL        1   5957.4 5991.4
## - intJOB          1   5960.3 5994.3
## - blnPARENT1      1   5962.2 5996.2
## - INCOME          1   5964.4 5998.4
## - intEDUCATION    1   5967.1 6001.1
## - BLUEBOOK        1   5976.1 6010.1
## - CLM_FREQ        1   5985.3 6019.3
## - intCAR_TYPE     1   5985.4 6019.4
## - KIDSDRIV        1   5987.0 6021.0
## - TIF             1   6007.4 6041.4
## - blnMSTATUS      1   6014.0 6048.0
## - MVR_PTS         1   6014.2 6048.2
## - TRAVTIME        1   6017.9 6051.9
## - blnNOT_REVOKED  1   6027.3 6061.3
## - blnCAR_USE      1   6130.3 6164.3
## - blnURBANICITY   1   6444.2 6478.2
##
## Step:  AIC=5990.16
## TARGET_FLAG ~ KIDSDRIV + INCOME + HOME_VAL + TRAVTIME + BLUEBOOK +
##     TIF + CLM_FREQ + MVR_PTS + blnPARENT1 + blnMSTATUS + blnCAR_USE +
##     blnNOT_REVOKED + blnURBANICITY + intEDUCATION + intJOB +
##     intCAR_TYPE
##
##                  Df Deviance    AIC
## <none>                5956.2 5990.2
## - HOME_VAL        1   5959.6 5991.6
## - intJOB          1   5962.8 5994.8
## - INCOME          1   5965.5 5997.5
## - intEDUCATION    1   5969.4 6001.4
## - blnPARENT1      1   5972.1 6004.1
## - BLUEBOOK        1   5978.2 6010.2
## - CLM_FREQ        1   5987.3 6019.3
## - intCAR_TYPE     1   5987.4 6019.4
## - KIDSDRIV        1   6001.8 6033.8
## - TIF             1   6009.0 6041.0
## - blnMSTATUS      1   6015.5 6047.5
## - MVR_PTS         1   6016.3 6048.3
## - TRAVTIME        1   6019.1 6051.1
## - blnNOT_REVOKED  1   6029.8 6061.8
## - blnCAR_USE      1   6131.8 6163.8
## - blnURBANICITY   1   6445.8 6477.8


##
```

```
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + INCOME + HOME_VAL + TRAVTIME +
##     BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + blnPARENT1 + blnMSTATUS +
##     blnCAR_USE + blnNOT_REVOKED + blnURBANICITY + intEDUCATION +
##     intJOB + intCAR_TYPE, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4243  -0.7277  -0.4205   0.6688   2.9483
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     3.721e+00  1.565e+00   2.377 0.017464 *
## KIDSDRIV        4.144e-01  6.113e-02   6.778 1.21e-11 ***
## INCOME         -3.867e-06  1.278e-06  -3.026 0.002481 **
## HOME_VAL       -2.328e-01  1.245e-01  -1.869 0.061586 .
## TRAVTIME        4.354e-01  5.640e-02   7.720 1.16e-14 ***
## BLUEBOOK       -2.799e-01  5.929e-02  -4.722 2.34e-06 ***
## TIF            -5.813e-02  8.153e-03  -7.129 1.01e-12 ***
## CLM_FREQ        1.582e-01  2.818e-02   5.614 1.97e-08 ***
## MVR_PTS         1.165e-01  1.508e-02   7.726 1.11e-14 ***
## blnPARENT1      4.148e-01  1.039e-01   3.992 6.56e-05 ***
## blnMSTATUS     -6.013e-01  7.778e-02  -7.730 1.07e-14 ***
## blnCAR_USE     -9.512e-01  7.257e-02 -13.107  < 2e-16 ***
## blnNOT_REVOKED -7.663e-01  8.856e-02  -8.653  < 2e-16 ***
## blnURBANICITY   2.290e+00  1.238e-01  18.496  < 2e-16 ***
## intEDUCATION   -1.388e-01  3.828e-02  -3.626 0.000288 ***
## intJOB         -6.990e-02  2.712e-02  -2.578 0.009946 **
## intCAR_TYPE    -1.488e-01  2.668e-02  -5.576 2.46e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7559.6  on 6527  degrees of freedom
## Residual deviance: 5956.2  on 6511  degrees of freedom
## AIC: 5990.2
##
## Number of Fisher Scoring iterations: 5
```




```
## [1] 0.7946423
```

```
##            Reference
## Prediction    0    1
##          0 1118  244
##          1   97  174
```

Our derived logistic regression model has a maximize AUC value of .79 with great p-values on all of the selected variables.

Below is table illustrating the various fitness parameters that describe the effectiveness of the logistic regression model. All the models are good - from a practical perspective, there is no difference between them.

Parameters	Model3
Accuracy	0.7911819
Classification Error Rate	0.2088181
Precision	0.8208517
Sensitivity	0.9201646
Specificity	0.4162679
F1 Score	0.8676756

Choose Model

Choose Model

We would like to pick Backward logistic regression model to make prediction for the evaluation dataset (non-zero value). This model has accuracy rate as high as 80%. In the meantime, the precision and sensitivity are at the level of 82% and 92%, which indicate this model is very good at eliminating false negative and false positive situations. Both AUC and F1 Score are around 80%, which is also telling us that it has high accuracy in terms of predicting the final response variables.

For linear regression model, all the models we created only contain one variable, which is BLUEBOOK. According to the following summary statistics, none of the model is performing significantly better than the others. They have very similar p-value, mean squared error, r-squared, and F-statistics. The R-sqaured is very low, even the best model we have is only 1.63%, which might indicate our model lack validity. Or it could also indicate the claim amount of motor vehicle accident tends to be unpredictable.

```
##           Parameters           m1           m2           m3           m4
## 1            p-value 1.537180e-07 2.630657e-48 2.226374e-08 1.351799e-08
## 2 Mean Squared Error 5.889033e+07 4.547081e+00 6.475595e-01 1.966487e+00
## 3                R^2 1.732972e-02 1.817743e-01 1.871589e-02 1.914363e-02
## 4       F-Statistics 4.726271e+00 5.953800e+01 5.111526e+00 5.230627e+00
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK, data = insi)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1570  -1508  -1493   -473 106091
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1790.48     785.63   2.279   0.0227 *
## BLUEBOOK      -30.14      82.56  -0.365   0.7151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4704 on 8159 degrees of freedom
## Multiple R-squared:  1.633e-05,  Adjusted R-squared:  -0.0001062
## F-statistic: 0.1332 on 1 and 8159 DF,  p-value: 0.7151
```

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIM
0	1489.168	0	48	0	11	52881	11.97756	3.2580
0	1493.656	1	40	1	11	50815	11.97756	3.0445
0	1528.792	0	44	2	12	43486	11.97756	3.4011
0	1515.304	0	35	2	0	21204	11.97756	4.3040
0	1499.837	0	59	0	12	87460	11.97756	3.8066
0	1484.489	0	46	0	14	61509	12.24298	1.9459
0	1509.233	0	60	0	12	37940	12.11581	2.7725
0	1486.504	0	54	0	12	33212	11.97308	3.2958
0	1482.732	2	36	2	12	130540	12.74896	1.6094

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIM
0	1475.874	0	50	0	8	167469	11.97756	3.0910
0	1487.203	0	42	0	13	52988	12.07980	3.1780
1	1501.733	0	41	2	7	17755	11.91046	3.3672
1	1514.883	1	37	2	13	59379	11.97756	4.1271
0	1510.850	0	36	3	12	56048	11.83382	2.7080
0	1501.299	0	34	3	12	22510	11.72713	3.2580
1	1513.715	0	35	2	12	39066	11.97756	3.7612
1	1570.066	2	44	2	14	45576	11.96160	3.2958
0	1487.487	0	48	0	9	61509	12.18300	3.6888
1	1515.501	0	62	0	15	40656	12.24646	4.0430
0	1510.431	0	39	0	11	33727	11.97756	3.2958

The Smooth Operators of R Fusion Have Struck Again.