# Homework 4

*Group 1*

## Contents



Prepared for:

Dr. Nathan Bastian

City University of New York, School of Professional Studies - Data 621


Prepared by:

Group 1

Senthil Dhanapal

Yadu Chittampalli

Christophe Hunt

# 1 Introduction

Consumers who own a car are often required to purchase car insurance to protect themselves from serious financial repercussions of being involved in a car accident. Insurance Providers must determine the risk of offering insurance coverage to a new customer through accurate statistical models that evaluate the consumers propensity for accidents. Since Insurance Providers are motivated by collecting the maximum amount of revenue from consumers while returning the lowest amount in accident claims, statistical modeling provides Insurance Providers with insight into the consumers behavior and the most appropriate pricing schemes[1].

# 2 Statement of the Problem

The purpose of this report is to develop statistical models to make inference into the likelihood of a customer being involved in a car accident and the cost associated of a customer being involved in a car accident.

# 3 Data Exploration

## 3.1 Variables Explained

The variables provided in the `Insurance Training Data Set` are explained below:

| Variable Code | Definition |
|---|---|
| INDEX | Identification Variable (do not use) |
| TARGET_FLAG | Was Car in a crash? 1=YES 0=NO |
| TARGET_AMT | If car was in a crash, what was the cost |
| AGE | Age of Driver |
| BLUEBOOK | Value of Vehicle |
| CAR_AGE | Vehicle Age |
| CAR_TYPE | Type of Car |
| CAR_USE | Vehicle Use |
| CLM_FREQ | # Claims (Past 5 Years) |
| EDUCATION | Max Education Level |
| HOMEKIDS | # Children at Home |
| HOME_VAL | Home Value |
| INCOME | Income |
| KIDSDRIV | # Driving Children |
| MSTATUS | Marital Status |
| MVR_PTS | Motor Vehicle Record Points |
| OLDCLAIM | Total Claims (Past 5 Years) |
| PARENT1 | Single Parent |
| RED_CAR | A Red Car |
| REVOKED | License Revoked (Past 7 Years) |
| SEX | Gender |
| TIF | Time in Force |
| TRAVTIME | Distance to Work |
| URBANICITY | Home/Work Area |
| YOJ | Years on Job |

[1] "Insider Information: How Insurance Companies Measure Risk - Insurance Companies.com." Insurance Companiescom. N.p., n.d. Web. 06 Nov. 2016.

### 3.1.1 Nominal Variables

We first look at our nominal variables and their applicable proportions. Interestingly, we see that in this data set only a quarter of the customer records indicate an accident occurred. Also, the majority of consumers in this data set have no kids at home, are married, more than a high school education but less than a PhD, use their car for private purposes, typically own a SUV or minivan, and also live in an urban environment. This provides an interesting insight to the type of customer this data set represents and should be considered when further interpreting our statistical model. Additionally, we should be mindful of any selection biases in this data set as consumers with extremely risky histories are likely to have not been extended insurance coverage.

Table 2: Table of nominal variables

| **Variable** | **Levels** | $\mathbf{n}$ | $\%$ | $\sum\%$ |
|---|---|---|---|---|
| TARGET_FLAG | 0 | 6008 | 73.6 | 73.6 |
| | 1 | 2153 | 26.4 | 100.0 |
| | all | 8161 | 100.0 | |
| KIDSDRIV | 0 | 7180 | 88.0 | 88.0 |
| | 1 | 636 | 7.8 | 95.8 |
| | 2 | 279 | 3.4 | 99.2 |
| | 3 | 62 | 0.8 | 100.0 |
| | 4 | 4 | 0.0 | 100.0 |
| | all | 8161 | 100.0 | |
| HOMEKIDS | 0 | 5289 | 64.8 | 64.8 |
| | 1 | 902 | 11.1 | 75.9 |
| | 2 | 1118 | 13.7 | 89.6 |
| | 3 | 674 | 8.3 | 97.8 |
| | 4 | 164 | 2.0 | 99.8 |
| | 5 | 14 | 0.2 | 100.0 |
| | all | 8161 | 100.0 | |
| PARENT1 | No | 7084 | 86.8 | 86.8 |
| | Yes | 1077 | 13.2 | 100.0 |
| | all | 8161 | 100.0 | |
| MSTATUS | No | 3267 | 40.0 | 40.0 |
| | Yes | 4894 | 60.0 | 100.0 |
| | all | 8161 | 100.0 | |
| SEX | F | 4375 | 53.6 | 53.6 |
| | M | 3786 | 46.4 | 100.0 |
| | all | 8161 | 100.0 | |
| EDUCATION | Less Than High School | 1203 | 14.7 | 14.7 |
| | High School | 2330 | 28.6 | 43.3 |
| | Bachelors | 2242 | 27.5 | 70.8 |
| | Masters | 1658 | 20.3 | 91.1 |
| | PhD | 728 | 8.9 | 100.0 |
| | all | 8161 | 100.0 | |
| JOB | | 526 | 6.4 | 6.4 |
| | Blue Collar | 1825 | 22.4 | 28.8 |
| | Clerical | 1271 | 15.6 | 44.4 |
| | Doctor | 246 | 3.0 | 47.4 |
| | Home Maker | 641 | 7.8 | 55.2 |
| | Lawyer | 835 | 10.2 | 65.5 |
| | Manager | 988 | 12.1 | 77.6 |
| | Professional | 1117 | 13.7 | 91.3 |
| | Student | 712 | 8.7 | 100.0 |
| | all | 8161 | 100.0 | |
| CAR_USE | Commercial | 3029 | 37.1 | 37.1 |
| | Private | 5132 | 62.9 | 100.0 |
| | all | 8161 | 100.0 | |
| CAR_TYPE | Minivan | 2145 | 26.3 | 26.3 |
| | Panel Truck | 676 | 8.3 | 34.6 |
| | Pickup | 1389 | 17.0 | 51.6 |
| | Sports Car | 907 | 11.1 | 62.7 |
| | SUV | 2294 | 28.1 | 90.8 |

Table 2: Table of nominal variables

| Variable | Levels | n | % | ∑% |
|---|---|---|---|---|
| | Van | 750 | 9.2 | 100.0 |
| | all | 8161 | 100.0 | |
| RED_CAR | no | 5783 | 70.9 | 70.9 |
| | yes | 2378 | 29.1 | 100.0 |
| | all | 8161 | 100.0 | |
| CLM_FREQ | 0 | 5009 | 61.4 | 61.4 |
| | 1 | 997 | 12.2 | 73.6 |
| | 2 | 1171 | 14.3 | 88.0 |
| | 3 | 776 | 9.5 | 97.5 |
| | 4 | 190 | 2.3 | 99.8 |
| | 5 | 18 | 0.2 | 100.0 |
| | all | 8161 | 100.0 | |
| REVOKED | No | 7161 | 87.8 | 87.8 |
| | Yes | 1000 | 12.2 | 100.0 |
| | all | 8161 | 100.0 | |
| URBANICITY | Highly Rural/ Rural | 1669 | 20.4 | 20.4 |
| | Highly Urban/ Urban | 6492 | 79.5 | 100.0 |
| | all | 8161 | 100.0 | |

### 3.1.2 Continuous and Discrete Variables

We can see that in our continuous and discrete variables there is some additional variability. The median claim amount (`TARGET_AMT`) is 0 which would coincide with only a quarter for records indicating an accident. However, the spread is large since the average payout is only $1,504.30 but the maximum payout was $107,586.10. Surprisingly, the median `AGE` is 45 and the average `AGE` is 44.8 years, while we expected a lower average it could be due to simple selection bias in the data set source or the aging US population bringing this average higher [2]. We also noticed that an `INCOME` of $0.00 seems unwise because it is unclear how the individual would be able to cover their premium costs without parental support. Finally, we should note that the data set has as `CAR_AGE` of -3, which is impossible and will need to be removed.

There are many missing values for this portion of our data set, we have over 400 values missing for years on the job, income, home value, and car age. Due to these missing values we will need to impute to complete our statistical model.

| Variable | n | Min | $q_1$ | $\tilde{x}$ | $\bar{x}$ | $q_3$ | Max | s | IQR | #NA |
|---|---|---|---|---|---|---|---|---|---|---|
| TARGET_AMT | 8161 | 0 | 0 | 0 | 1504.3 | 1036 | 107586.1 | 4704.0 | 1036 | 0 |
| TIF | 8161 | 1 | 1 | 4 | 5.4 | 7 | 25.0 | 4.1 | 6 | 0 |
| AGE | 8155 | 16 | 39 | 45 | 44.8 | 51 | 81.0 | 8.6 | 12 | 6 |
| YOJ | 7707 | 0 | 9 | 11 | 10.5 | 13 | 23.0 | 4.1 | 4 | 454 |
| INCOME | 7716 | 0 | 28097 | 54028 | 61898.1 | 85986 | 367030.0 | 47572.7 | 57889 | 445 |
| HOME_VAL | 7697 | 0 | 0 | 161160 | 154867.3 | 238724 | 885282.0 | 129123.8 | 238724 | 464 |
| TRAVTIME | 8161 | 5 | 22 | 33 | 33.5 | 44 | 142.0 | 15.9 | 22 | 0 |
| BLUEBOOK | 8161 | 1500 | 9280 | 14440 | 15709.9 | 20850 | 69740.0 | 8419.7 | 11570 | 0 |
| OLDCLAIM | 8161 | 0 | 0 | 0 | 4037.1 | 4636 | 57037.0 | 8777.1 | 4636 | 0 |
| MVR_PTS | 8161 | 0 | 0 | 1 | 1.7 | 3 | 13.0 | 2.1 | 3 | 0 |
| CAR_AGE | 7651 | -3 | 1 | 8 | 8.3 | 12 | 28.0 | 5.7 | 11 | 510 |

Table 3:

[2] Ortman, Jennifer M., Victoria A. Velkoff, and Howard Hogan. "An aging nation: the older population in the United States." Washington, DC: US Census Bureau (2014): 25-1140.

## 3.2 Imputting Missing Values

In order to address the missing values in our variables we used a non-parametric imputation method (Random Forest) using the `missForest` package. The function is particularly useful in that it can handle any type of input data and it will make as few assumptions about the structure of the data as possible.[3]

**Table 2 : Imputed Descriptive Statistics**
**25 Variables 8161 Observations**

**TARGET_FLAG**

| n | missing | distinct | Info | Sum | Mean | Gmd |
|---|---|---|---|---|---|---|
| 8161 | 0 | 2 | 0.583 | 2153 | 0.3 | 0.4 |

**TARGET_AMT**

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 1949 | 0.601 | 1504 | 2574 | 0 | 0 | 0 | 0 | 1036 | 4904 | 6452 |

```
lowest :      0.00000     30.27728     58.53106     95.56732    108.74150
highest:  73783.46592  77907.43028  78874.19056  85523.65335 107586.13616
```

**KIDSDRIV**

| n | missing | distinct | Info | Mean | Gmd |
|---|---|---|---|---|---|
| 8161 | 0 | 5 | 0.318 | 0.2 | 0.3 |

```
lowest : 0 1 2 3 4, highest: 0 1 2 3 4
```

```
0 (7180, 0.880), 1 (636, 0.078), 2 (279, 0.034), 3 (62, 0.008), 4 (4, 0.000)
```

**AGE**

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 66 | 0.999 | 45 | 10 | 30 | 33 | 39 | 45 | 51 | 56 | 59 |

```
lowest : 16 17 18 19 20, highest: 72 73 76 80 81
```

**HOMEKIDS**

| n | missing | distinct | Info | Mean | Gmd |
|---|---|---|---|---|---|
| 8161 | 0 | 6 | 0.723 | 0.7 | 1 |

```
lowest : 0 1 2 3 4, highest: 1 2 3 4 5
```

```
0 (5289, 0.648), 1 (902, 0.111), 2 (1118, 0.137), 3 (674, 0.083), 4 (164, 0.020), 5 (14,
0.002)
```

**YOJ**

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 446 | 0.991 | 10 | 4 | 0 | 5 | 9 | 11 | 13 | 14 | 15 |

```
lowest :  0.00  0.15  0.20  0.26  0.27, highest: 16.00 17.00 18.00 19.00 23.00
```

**INCOME**

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 7057 | 1 | 61569 | 50845 | 0e+00 | 5e+03 | 3e+04 | 5e+04 | 9e+04 | 1e+05 | 2e+05 |

```
lowest :      0.00      5.00      7.00     18.00     26.33
highest: 306277.00 309628.00 320127.00 332339.00 367030.00
```

**PARENT1**

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 2 |

```
No (7084, 0.868), Yes (1077, 0.132)
```

**HOME_VAL**

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 5570 | 0.978 | 2e+05 | 1e+05 | 0e+00 | 0e+00 | 0e+00 | 2e+05 | 2e+05 | 3e+05 | 4e+05 |

```
lowest :      0.000   4176.960   8196.080   8263.335   9438.009
highest: 657804.000 682634.000 738153.000 750455.000 885282.000
```

[3] Stekhoven, Daniel J., and Peter Bühlmann. "MissForest-non-parametric missing value imputation for mixed-type data." Bioinformatics 28.1 (2012): 112-118.

### MSTATUS

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 2 |

No (3267, 0.4), Yes (4894, 0.6)

### SEX

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 2 |

F (4375, 0.536), M (3786, 0.464)

### EDUCATION

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 5 |

lowest : Bachelors High School Less Than High School Masters PhD
highest: Bachelors High School Less Than High School Masters PhD

Bachelors (2242, 0.275), High School (2330, 0.286), Less Than High School (1203, 0.147), Masters (1658, 0.203), PhD (728, 0.089)

### JOB

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 8 |

lowest : Blue Collar Clerical Doctor Home Maker Lawyer
highest: Home Maker Lawyer Manager Professional Student

| Value | Blue Collar | Clerical | Doctor | Home Maker | Lawyer | Manager |
|---|---|---|---|---|---|---|
| Frequency | 1830 | 1273 | 254 | 643 | 865 | 1412 |
| Proportion | 0.224 | 0.156 | 0.031 | 0.079 | 0.106 | 0.173 |

| Value | Professional | Student |
|---|---|---|
| Frequency | 1172 | 712 |
| Proportion | 0.144 | 0.087 |

### TRAVTIME

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 97 | 1 | 33 | 18 | 7 | 13 | 22 | 33 | 44 | 54 | 60 |

lowest : 5 6 7 8 9, highest: 103 113 124 134 142

### CAR_USE

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 2 |

Commercial (3029, 0.371), Private (5132, 0.629)

### BLUEBOOK

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 2789 | 1 | 15710 | 9354 | 4900 | 6000 | 9280 | 14440 | 20850 | 27460 | 31110 |

lowest : 1500 1520 1530 1540 1590, highest: 57970 61050 62240 65970 69740

### TIF

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8161 | 0 | 23 | 0.961 | 5 | 5 | 1 | 1 | 1 | 4 | 7 | 11 | 13 |

lowest : 1 2 3 4 5, highest: 19 20 21 22 25

### CAR_TYPE

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 6 |

lowest : Minivan Panel Truck Pickup Sports Car SUV
highest: Panel Truck Pickup Sports Car SUV Van

Minivan (2145, 0.263), Panel Truck (676, 0.083), Pickup (1389, 0.170), Sports Car (907, 0.111), SUV (2294, 0.281), Van (750, 0.092)

### RED_CAR

| n | missing | distinct |
|---|---|---|
| 8161 | 0 | 2 |

```
no (5783, 0.709), yes (2378, 0.291)
```

**OLDCLAIM**

```
     n    missing   distinct       Info       Mean        Gmd        .05        .10        .25        .50        .75        .90        .95
  8161          0       2857      0.769       4037       6563          0          0          0          0       4636       9583      27090

lowest :     0   502   506   518   519, highest: 52507 53477 53568 53986 57037
```

**CLM_FREQ**

```
     n    missing   distinct       Info       Mean        Gmd
  8161          0          6      0.763        0.8          1

lowest : 0 1 2 3 4, highest: 1 2 3 4 5

0 (5009, 0.614), 1 (997, 0.122), 2 (1171, 0.143), 3 (776, 0.095), 4 (190, 0.023), 5 (18,
0.002)
```

**REVOKED**

```
     n    missing   distinct
  8161          0          2

No (7161, 0.877), Yes (1000, 0.123)
```

**MVR_PTS**

```
     n    missing   distinct       Info       Mean        Gmd        .05        .10        .25        .50        .75        .90        .95
  8161          0         13        0.9          2          2          0          0          0          1          3          5          6

lowest :  0  1  2  3  4, highest:  8  9 10 11 13

Value          0     1     2     3     4     5     6     7     8     9    10    11    13
Frequency   3712  1157   948   758   599   399   266   167    84    45    13    11     2
Proportion 0.455 0.142 0.116 0.093 0.073 0.049 0.033 0.020 0.010 0.006 0.002 0.001 0.000
```

**CAR_AGE**

```
     n    missing   distinct       Info       Mean        Gmd        .05        .10        .25        .50        .75        .90        .95
  8161          0        507      0.985          8          6          1          1          4          8         12         16         18

lowest :  0.000  1.000  2.000  2.035  2.890, highest: 24.000 25.000 26.000 27.000 28.000
```

**URBANICITY**

```
     n    missing   distinct
  8161          0          2

Highly Rural/ Rural (1669, 0.205), Highly Urban/ Urban (6492, 0.795)
```

## 3.3 Exploration of Variables

# 4 Data Transformation

## 4.1 Outliers Treatment

We chose winsorizing as the method to address outliers. Instead of trimming values, winsorizing uses the interquantile range to replace values that are above or below the interquantile range multiplied by a factor. Those values above or below the range multiplied by the factor are then replaced with max and min value of the interquantile range. Using the factor 2.2 for winsorizing outliers is a method developed my Hoaglin and Iglewicz and published Journal of American Statistical Association in 1987[4].

The below table is the summary results of the winsorizing of the data.

[4]Hoaglin, D. C., and Iglewicz, B. (1987), Fine tuning some resistant rules for outlier labeling, Journal of American Statistical Association, 82, 1147-1149.

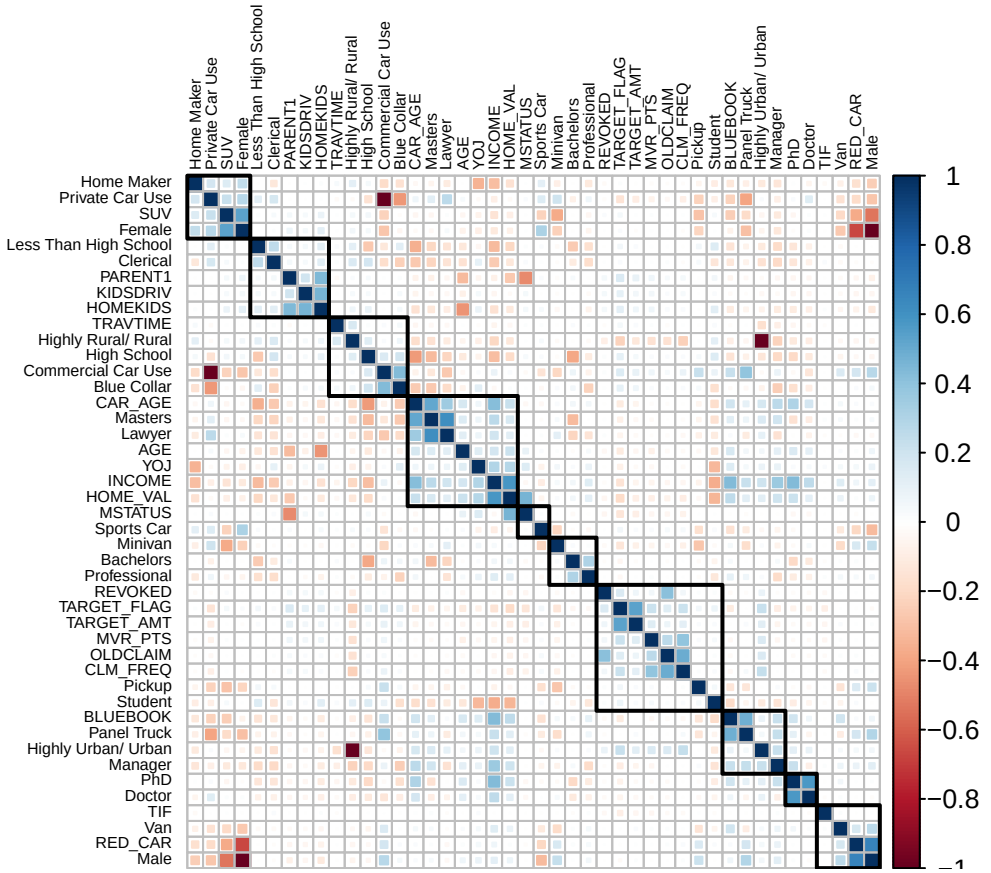



Figure 1: Correlation Plot of Training Data Set

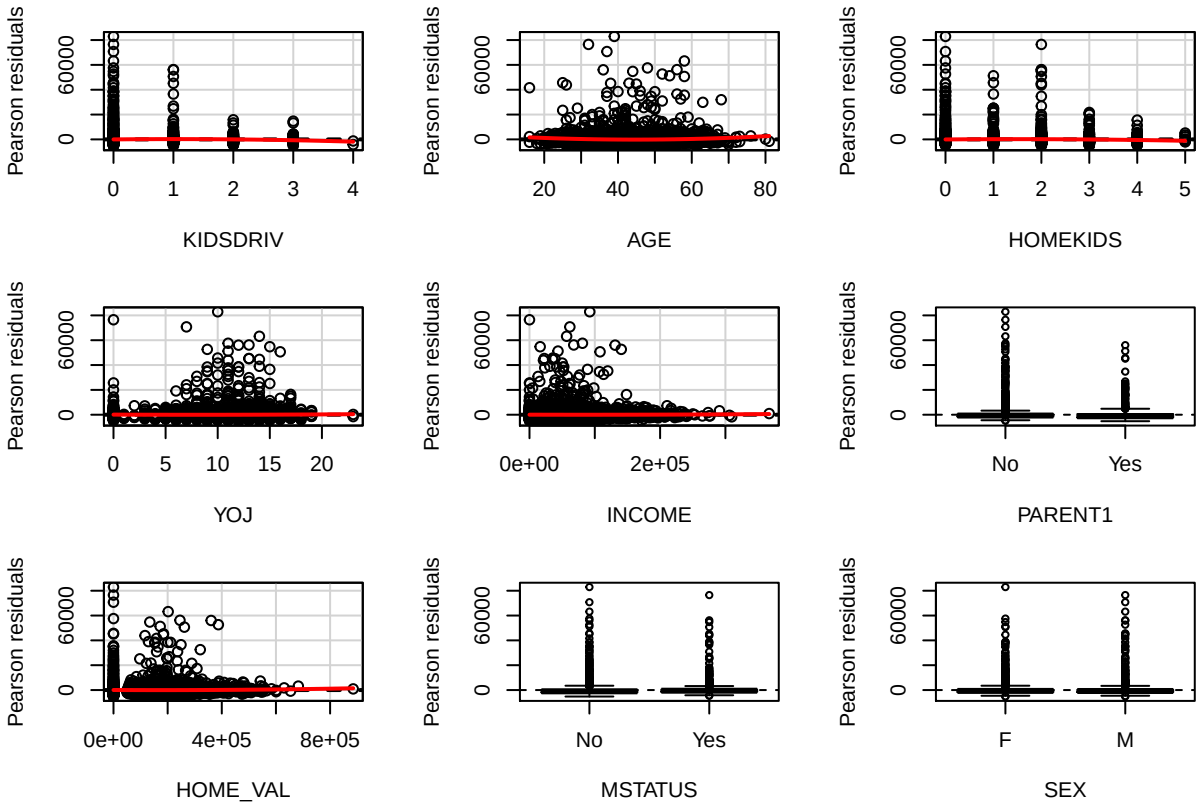
Table 4:

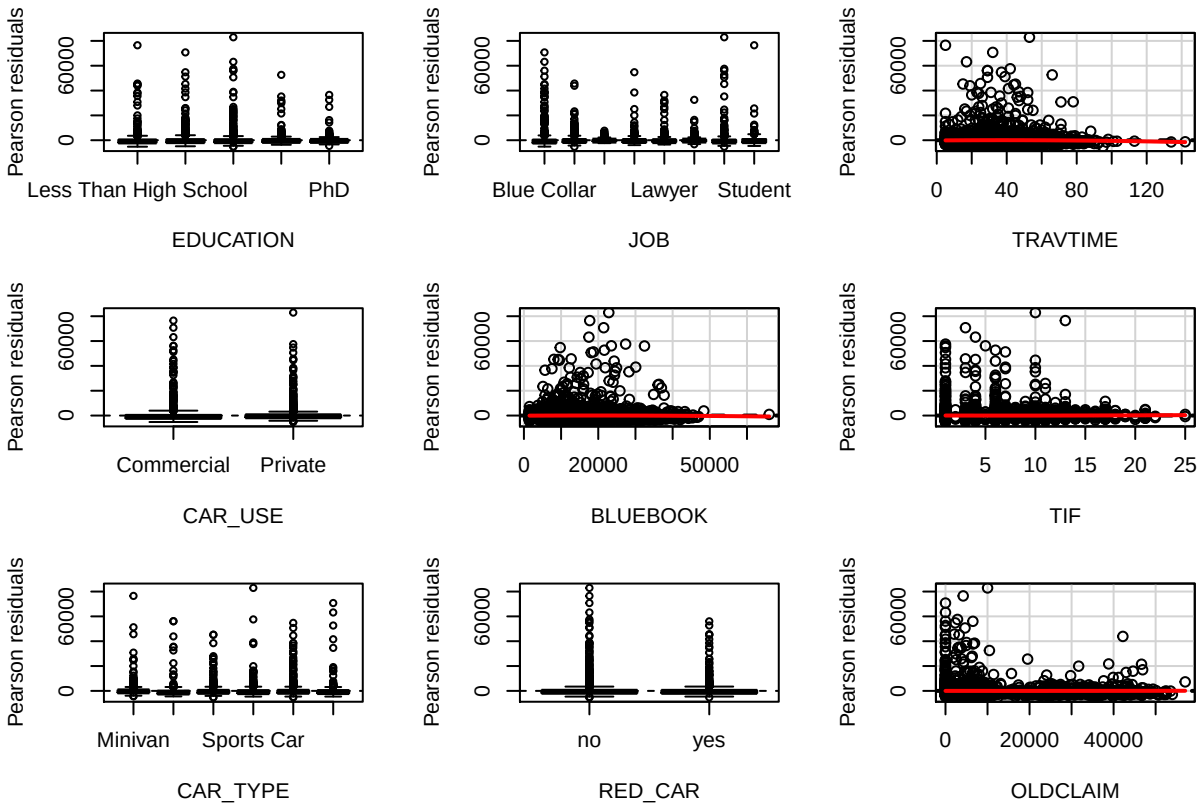
| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| INCOME | 8,161 | 61,225.430 | 45,828.110 | 0.000 | 210,660.000 |
| HOME_VAL | 8,161 | 154,982.200 | 127,401.100 | 0.000 | 750,455.000 |
| OLDCLAIM | 8,161 | 2,757.246 | 4,469.462 | 0 | 14,738 |
| BLUEBOOK | 8,161 | 15,696.740 | 8,359.971 | 1,500 | 46,250 |
| TARGET_FLAG | 8,161 | 0.264 | 0.441 | 0 | 1 |
| TARGET_AMT | 8,161 | 1,504.325 | 4,704.027 | 0.000 | 107,586.100 |
| KIDSDRIV | 8,161 | 0.171 | 0.512 | 0 | 4 |
| AGE | 8,161 | 44.782 | 8.630 | 16.000 | 81.000 |
| HOMEKIDS | 8,161 | 0.721 | 1.116 | 0 | 5 |
| YOJ | 8,161 | 10.498 | 4.037 | 0.000 | 23.000 |
| TRAVTIME | 8,161 | 33.486 | 15.908 | 5 | 142 |
| TIF | 8,161 | 5.351 | 4.147 | 1 | 25 |
| CLM_FREQ | 8,161 | 0.799 | 1.158 | 0 | 5 |
| MVR_PTS | 8,161 | 1.696 | 2.147 | 0 | 13 |
| CAR_AGE | 8,161 | 8.350 | 5.602 | 0.000 | 28.000 |

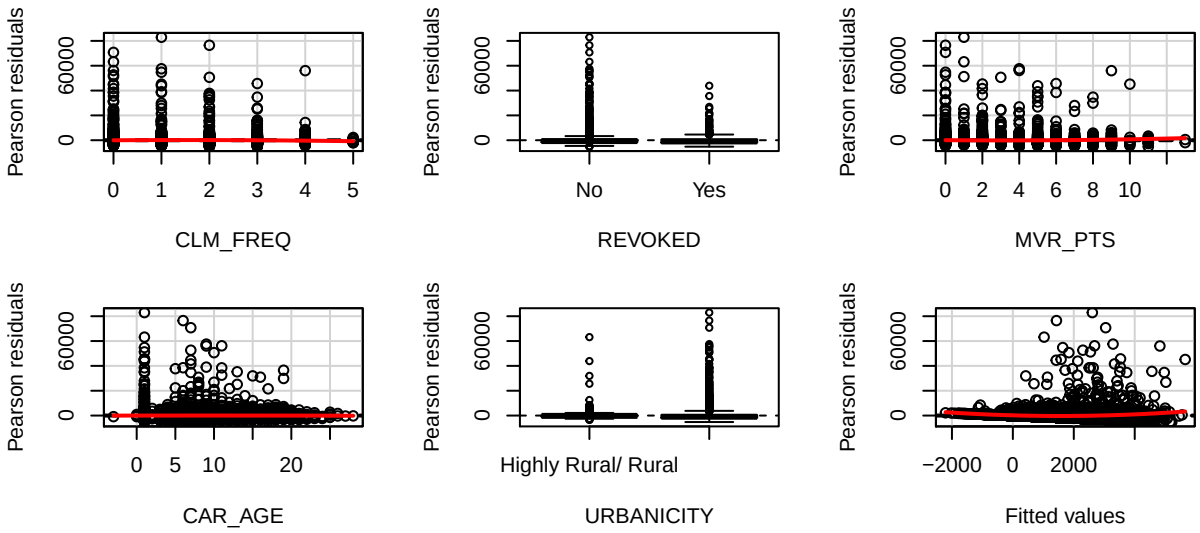
## 4.2 BoxCox Transformations

The Box-Cox transformations were done only on three of the input variables - income, house value, and the total number of claims during the past 5 years. These transformations were done based on the residual plots. In the residual plots, these three variables showed a great deal of non-constant variance because the plots were funnel-shaped.

```
## Non-constant Variance Score Test
## Variance formula: ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 + HOME_VAL + MSTATUS + SEX + I
## Chisquare = 5720.739    Df = 36     p = 0
```




Pearson residuals
0
60000
Less Than High School
PhD
EDUCATION
Blue Collar
Lawyer
Student
JOB
0
40
80
120
TRAVTIME
Commercial
Private
CAR_USE
0
20000
50000
BLUEBOOK
5
10
15
20
25
TIF
Minivan
Sports Car
CAR_TYPE
no
yes
RED_CAR
0
20000
40000
OLDCLAIM




```
##            Test stat Pr(>|t|)
## KIDSDRIV      -1.733    0.083
## AGE            4.364    0.000
## HOMEKIDS      -2.420    0.016
## YOJ            1.003    0.316
## INCOME         0.513    0.608
## PARENT1           NA       NA
## HOME_VAL       1.483    0.138
## MSTATUS           NA       NA
## SEX               NA       NA
## EDUCATION         NA       NA
## JOB               NA       NA
## TRAVTIME      -1.033    0.302
## CAR_USE           NA       NA
## BLUEBOOK      -0.618    0.536
## TIF            0.467    0.640
## CAR_TYPE          NA       NA
## RED_CAR           NA       NA
## OLDCLAIM       0.171    0.864
## CLM_FREQ      -1.784    0.074
## REVOKED           NA       NA
## MVR_PTS        1.987    0.047
## CAR_AGE       -0.277    0.782
## URBANICITY        NA       NA
## Tukey test     8.214    0.000
```

Using the `BoxCox.lambda` function from the `forecast` package we are able to determine our necessary

transformations to our independent variables.

| $\lambda$ | Variables |
|---|---|
| 0.268842617694589 | INCOME |
| 0.505233636014921 | HOME_VAL |
| 0.456635555660553 | TRAVTIME |

Utilizing the below table of common transformations based on the lambda value of the BoxCox we further transform our independent variables.

Common Box-Cox Transformations[5] [6]

| $\lambda$ | Y' |
|---|---|
| -2 | $Y^{-2} = \frac{1}{Y^2}$ |
| -1 | $Y^{-1} = \frac{1}{Y^1}$ |
| -0.5 | $Y^{-0.5} = \frac{1}{\sqrt{(Y)}}$ |
| 0 | $\log(Y)$ |
| .25 | $\sqrt[4]{Y}$ |
| 0.5 | $Y^{0.5} = \sqrt{(Y)}$ |
| 1 | $Y^1 = Y$ |
| 2 | $Y^2$ |

Lambda values were truncated to the nearest tenth that match a common transformation as per the below table.

| variable | variable transformation |
|---|---|
| INCOME | $\sqrt[4]{INCOME}$ |
| HOME VAL | $\sqrt{(HOME\ VAL)}$ |
| TRAVTIME | $\sqrt{TRAVTIME}$ |

[5] By Understanding Both the Concept of Transformation and the Box-Cox Method, Practitioners Will Be Better Prepared to Work with Non-normal Data. . "Making Data Normal Using Box-Cox Power Transformation." ISixSigma. N.p., n.d. Web. 29 Oct. 2016.

[6] Osborne, Jason W. "Improving your data transformations: Applying the Box-Cox transformation." Practical Assessment, Research & Evaluation 15.12 (2010): 1-9.

# 5 Models Built

## 5.1 Logistic Model 1 - Backwards Selection Method

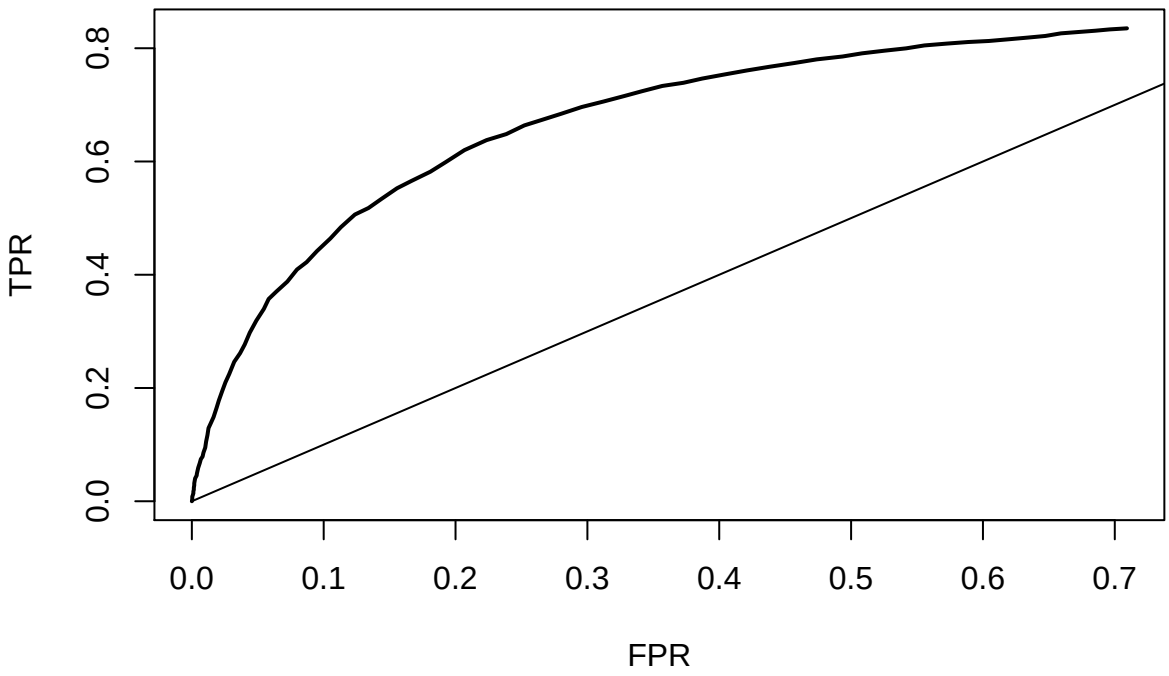
In the backward step selection model The resulting AIC was .

Table 6:

| | fullModel |
|---|---|
| Constant | 0.113*** (0.037) |
| KIDSDRIV | 0.065*** (0.009) |
| I(INCOMEˆ(1/4)) | −0.008*** (0.002) |
| PARENT1Yes | 0.085*** (0.015) |
| I(sqrt(HOME_VAL)) | −0.0001*** (0.00003) |
| MSTATUSYes | −0.059*** (0.012) |
| EDUCATIONHigh School | 0.062*** (0.013) |
| EDUCATIONLess Than High School | 0.054*** (0.016) |
| EDUCATIONMasters | 0.024 (0.017) |
| EDUCATIONPhD | 0.041* (0.022) |
| JOBClerical | 0.012 (0.016) |
| JOBDoctor | −0.134*** (0.036) |
| JOBHome Maker | −0.053** (0.025) |
| JOBLawyer | −0.051* (0.026) |
| JOBManager | −0.136*** (0.019) |
| JOBProfessional | −0.028 (0.018) |
| JOBStudent | −0.069*** (0.023) |
| I(sqrt(TRAVTIME)) | 0.023*** (0.003) |
| CAR_USEPrivate | −0.126*** (0.014) |
| BLUEBOOK | −0.00000*** (0.00000) |
| TIF | −0.008*** (0.001) |
| CAR_TYPEPanel Truck | 0.069*** (0.022) |
| CAR_TYPEPickup | 0.073*** (0.015) |
| CAR_TYPESports Car | 0.133*** (0.016) |
| CAR_TYPESUV | 0.096*** (0.012) |
| CAR_TYPEVan | 0.081*** (0.018) |
| OLDCLAIM | −0.00000** (0.00000) |
| CLM_FREQ | 0.031*** (0.005) |
| REVOKEDYes | 0.139*** (0.014) |
| MVR_PTS | 0.021*** (0.002) |
| URBANICITYHighly Urban/ Urban | 0.299*** (0.012) |
| *N* | 8,161 |
| Log Likelihood | −3,820.565 |
| Akaike Inf. Crit. | 7,703.130 |
| *Notes:* | ***Significant at the 1 percent level. |
| | **Significant at the 5 percent level. |
| | *Significant at the 10 percent level. |

### 5.1.1 Model Metrics for Backwards Selection

We first use an established threshold of .50 to determine our best possible threshold.



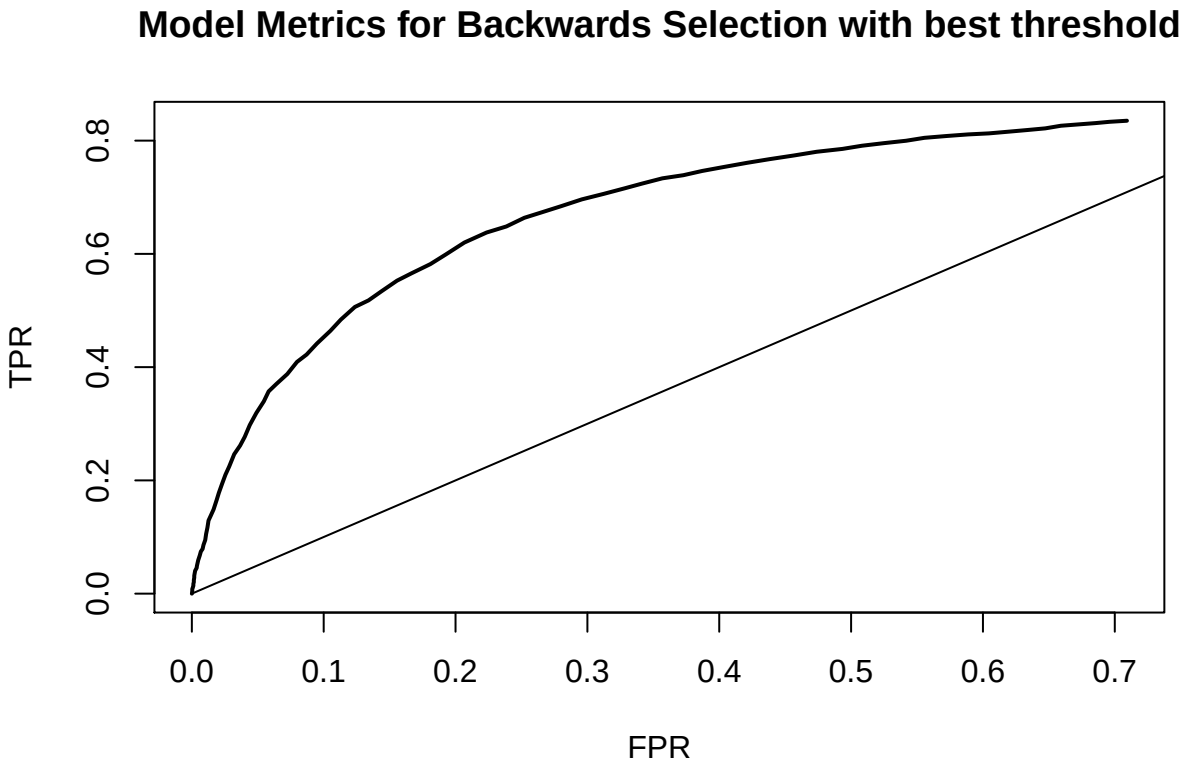

| | Act-Pos | Act-Neg |
|---|---|---|
| Pred-Pos | 687 | 1466 |
| Pred-Neg | 294 | 5714 |

| | Model Metrics for Backwards Selection |
|---|---|
| accuracy | 0.784 |
| classif.error | 0.216 |
| precision | 0.700 |
| sensitivity | 0.319 |
| specificity | 0.951 |
| f1score | 0.438 |
| auc | 0.469 |
| best.threshold | 0.320 |
| aic | 7703.130 |
| CVError | NA |

Our previous results indicate that .320 would be the best threshold for this model so we re-run our metrics using this threshold.

#### 5.1.1.1 Model Metrics for Backwards Selection with best threshold

Model Metrics using best threshold of .320.




| | Act-Pos | Act-Neg |
|---|---|---|
| Pred-Pos | 1373 | 780 |
| Pred-Neg | 1343 | 4665 |

| | Model Metrics for Backwards Selection with best threshold |
|---|---|
| accuracy | 0.740 |
| classif.error | 0.260 |
| precision | 0.506 |
| sensitivity | 0.638 |
| specificity | 0.776 |
| f1score | 0.564 |
| auc | 0.469 |
| best.threshold | 0.320 |
| aic | 7703.130 |
| CVError | NA |

### 5.1.2 Multicollinearity for Backwards Selection

| rn | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| KIDSDRIV | 1.077269 | 1 | 1.037916 |
| I(INCOME^(1/4)) | 3.303477 | 1 | 1.817547 |
| PARENT1 | 1.411040 | 1 | 1.187872 |
| I(sqrt(HOME_VAL)) | 2.093420 | 1 | 1.446866 |
| MSTATUS | 2.012248 | 1 | 1.418537 |
| EDUCATION | 6.363135 | 4 | 1.260256 |
| JOB | 25.141215 | 7 | 1.259005 |
| TRAVTIME | 1.037090 | 1 | 1.018376 |
| CAR_USE | 2.409461 | 1 | 1.552244 |
| BLUEBOOK | 1.671917 | 1 | 1.293026 |
| TIF | 1.007017 | 1 | 1.003502 |
| CAR_TYPE | 2.483562 | 5 | 1.095236 |
| CLM_FREQ | 1.262974 | 1 | 1.123821 |
| REVOKED | 1.020286 | 1 | 1.010092 |
| MVR_PTS | 1.213958 | 1 | 1.101798 |
| URBANICITY | 1.250699 | 1 | 1.118347 |

## 5.2 Logistic Model 2 - Forwards Selection Method

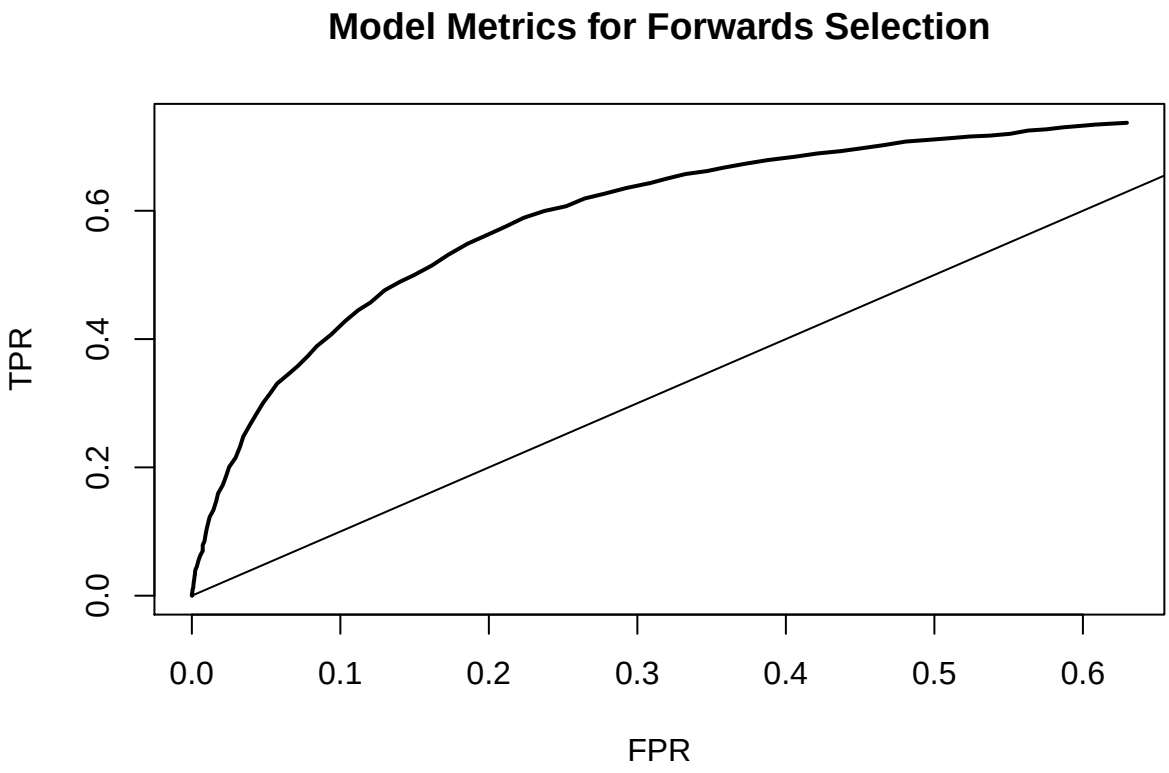
Table 12:

| | fullModel |
|---|---|
| Constant | 0.123** (0.048) |
| KIDSDRIV | 0.062*** (0.010) |
| AGE | −0.0004 (0.001) |
| HOMEKIDS | 0.004 (0.006) |
| YOJ | 0.001 (0.002) |
| I(INCOMEˆ(1/4)) | −0.009*** (0.002) |
| PARENT1Yes | 0.075*** (0.017) |
| I(sqrt(HOME_VAL)) | −0.0001*** (0.00003) |
| MSTATUSYes | −0.063*** (0.013) |
| SEXM | 0.014 (0.016) |
| EDUCATIONHigh School | 0.059*** (0.014) |
| EDUCATIONLess Than High School | 0.049*** (0.018) |
| EDUCATIONMasters | 0.028 (0.018) |
| EDUCATIONPhD | 0.046** (0.023) |
| JOBClerical | 0.010 (0.016) |
| JOBDoctor | −0.132*** (0.037) |
| JOBHome Maker | −0.051** (0.025) |
| JOBLawyer | −0.049* (0.026) |
| JOBManager | −0.134*** (0.019) |
| JOBProfessional | −0.027 (0.018) |
| JOBStudent | −0.070*** (0.023) |
| I(sqrt(TRAVTIME)) | 0.023*** (0.003) |
| CAR_USEPrivate | −0.126*** (0.014) |
| BLUEBOOK | −0.00000*** (0.00000) |
| TIF | −0.008*** (0.001) |
| CAR_TYPEPanel Truck | 0.063*** (0.024) |
| CAR_TYPEPickup | 0.074*** (0.015) |
| CAR_TYPESports Car | 0.141*** (0.019) |
| CAR_TYPESUV | 0.104*** (0.015) |
| CAR_TYPEVan | 0.077*** (0.018) |
| RED_CARyes | −0.005 (0.013) |
| OLDCLAIM | −0.00000** (0.00000) |
| CLM_FREQ | 0.031*** (0.005) |
| REVOKEDYes | 0.138*** (0.014) |
| MVR_PTS | 0.021*** (0.002) |
| CAR_AGE | −0.001 (0.001) |
| URBANICITYHighly Urban/ Urban | 0.299*** (0.012) |
| *N* | 8,161 |
| Log Likelihood | −3,819.035 |
| Akaike Inf. Crit. | 7,712.070 |

*Notes:* ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

### 5.2.1 Model Metrics for Forwards Selection

We first use an established threshold of .50 to determine our best possible threshold.



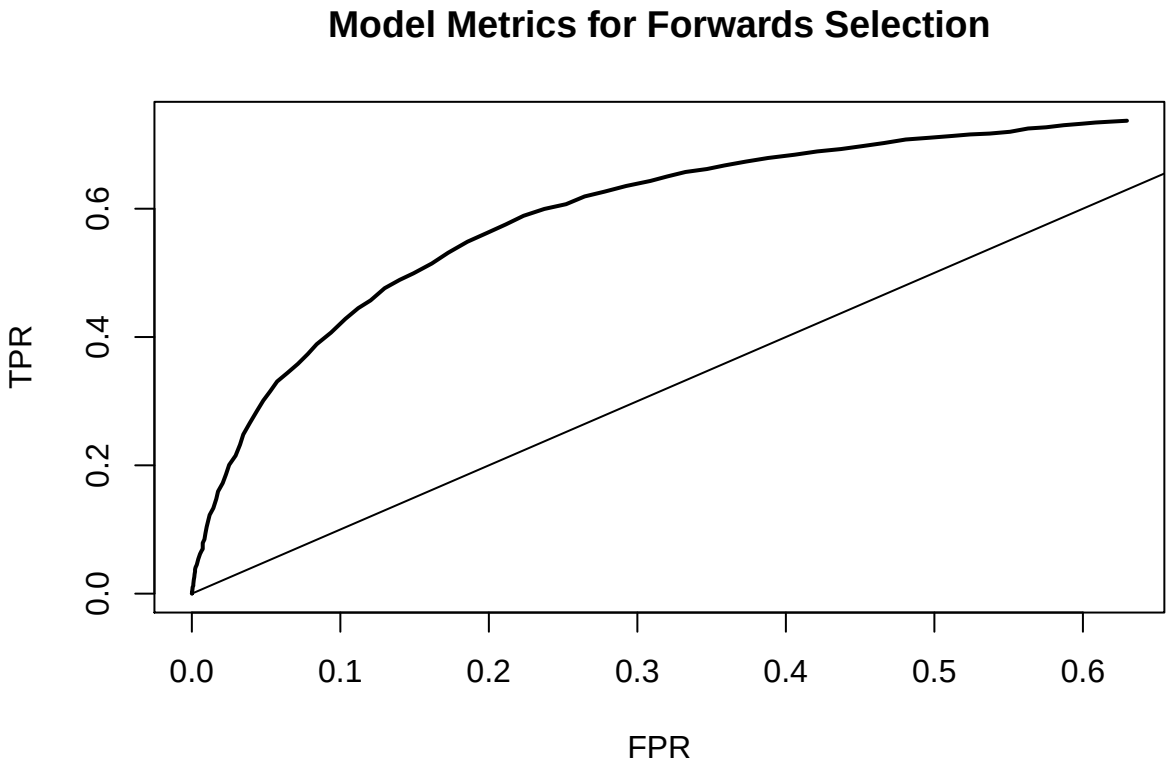

| | Act-Pos | Act-Neg |
|---|---|---|
| Pred-Pos | 614 | 1539 |
| Pred-Neg | 264 | 5744 |

| | Model Metrics for Forwards Selection |
|---|---|
| accuracy | 0.779 |
| classif.error | 0.221 |
| precision | 0.699 |
| sensitivity | 0.285 |
| specificity | 0.956 |
| f1score | 0.405 |
| auc | 0.368 |
| best.threshold | 0.300 |
| aic | 7712.070 |
| CVError | NA |

Our previous results indicate that .300 would be the best threshold for this model so we re-run our metrics using this threshold.

#### 5.2.1.1 Model Metrics for Forwards Selection with best threshold

Model Metrics using best threshold of .300.




| | Act-Pos | Act-Neg |
|---|---|---|
| Pred-Pos | 1268 | 885 |
| Pred-Neg | 1342 | 4666 |

| | Model Metrics for Forwards Selection |
|---|---|
| accuracy | 0.727 |
| classif.error | 0.273 |
| precision | 0.486 |
| sensitivity | 0.589 |
| specificity | 0.777 |
| f1score | 0.532 |
| auc | 0.368 |
| best.threshold | 0.300 |
| aic | 7712.070 |
| CVError | NA |

### 5.2.2 Multicollinearity for Forwards Selection

Here in our forward selection model we find that no variable exceeds our pre-established threshold of 5 for multicollinearity.

| rn | GVIF | Df | GVIFˆ(1/(2*Df)) |
|---|---|---|---|
| KIDSDRIV | 1.325344 | 1 | 1.151236 |
| AGE | 1.487782 | 1 | 1.219747 |
| HOMEKIDS | 2.136050 | 1 | 1.461523 |
| YOJ | 2.043591 | 1 | 1.429542 |
| I(INCOMEˆ(1/4)) | 4.626874 | 1 | 2.151017 |
| PARENT1 | 1.844294 | 1 | 1.358048 |
| I(sqrt(HOME_VAL)) | 2.115696 | 1 | 1.454543 |
| MSTATUS | 2.204211 | 1 | 1.484659 |
| SEX | 3.324170 | 1 | 1.823231 |
| EDUCATION | 8.969420 | 4 | 1.315514 |
| JOB | 26.685061 | 7 | 1.264376 |
| I(sqrt(TRAVTIME)) | 1.033912 | 1 | 1.016814 |
| CAR_USE | 2.409933 | 1 | 1.552396 |
| BLUEBOOK | 2.065326 | 1 | 1.437124 |
| TIF | 1.007641 | 1 | 1.003813 |
| CAR_TYPE | 5.489500 | 5 | 1.185641 |
| RED_CAR | 1.814257 | 1 | 1.346944 |
| OLDCLAIM | 2.193395 | 1 | 1.481011 |
| CLM_FREQ | 2.076612 | 1 | 1.441045 |
| REVOKED | 1.200779 | 1 | 1.095800 |
| MVR_PTS | 1.239981 | 1 | 1.113544 |
| CAR_AGE | 2.166209 | 1 | 1.471805 |
| URBANICITY | 1.248323 | 1 | 1.117284 |

## 5.3 Logistic Model 3 - Subset Selection Method

### 5.3.1 Subset Variable Selection

Using the `leaps` package and the `regsubsets` function we are able to subset our independent variables by looking at the best model for each predictor.

| | 1 ( 1 ) | 2 ( 1 ) | 3 ( 1 ) | 4 ( 1 ) | 5 ( 1 ) | 6 ( 1 ) | 7 ( 1 ) | 8 ( 1 ) |
|---|---|---|---|---|---|---|---|---|
| KIDSDRIV | | | | | | | | * |
| AGE | | | | | | | | |
| HOMEKIDS | | | | | | | | |
| YOJ | | | | | | | | |
| I(INCOMEˆ(1/4)) | | | | | * | * | * | * |
| PARENT1Yes | | | | | * | * | * | |
| I(sqrt(HOME_VAL)) | | * | * | * | | | | |
| MSTATUSYes | | | | | | | | * |
| SEXM | | | | | | | | |
| EDUCATIONHigh School | | | | | | | | |
| EDUCATIONLess Than High School | | | | | | | | |
| EDUCATIONMasters | | | | | | | | |
| EDUCATIONPhD | | | | | | | | |
| JOBClerical | | | | | | | | |
| JOBDoctor | | | | | | | | |
| JOBHome Maker | | | | | | | | |
| JOBLawyer | | | | | | | | |
| JOBManager | | | | | | | * | * |
| JOBProfessional | | | | | | | | |
| JOBStudent | | | | | | | | |
| I(sqrt(TRAVTIME)) | | | | | | | | |
| CAR_USEPrivate | | | | * | * | * | * | * |
| BLUEBOOK | | | | | | | | |
| TIF | | | | | | | | |
| CAR_TYPEPanel Truck | | | | | | | | |
| CAR_TYPEPickup | | | | | | | | |
| CAR_TYPESports Car | | | | | | | | |
| CAR_TYPESUV | | | | | | | | |
| CAR_TYPEVan | | | | | | | | |
| RED_CARyes | | | | | | | | |
| OLDCLAIM | | | | | | | | |
| CLM_FREQ | | | | | | | | |
| REVOKEDYes | | | | | | * | * | * |
| MVR_PTS | | | * | * | * | * | * | * |
| CAR_AGE | | | | | | | | |
| URBANICITYHighly Urban/ Urban | * | * | * | * | * | * | * | * |

### 5.3.2 Subset Model

The variables as indicated in column 8 of the previous table will be further implement into our subset selection model in the following table. We don't see as strong of a relationship in our independent variables to the dependent variable in this model as our previous model. For example, the coefficient for tax was as high as 106 in the forward selection model but it is -5 in this model. However, our intercept in this model is larger than any other model.

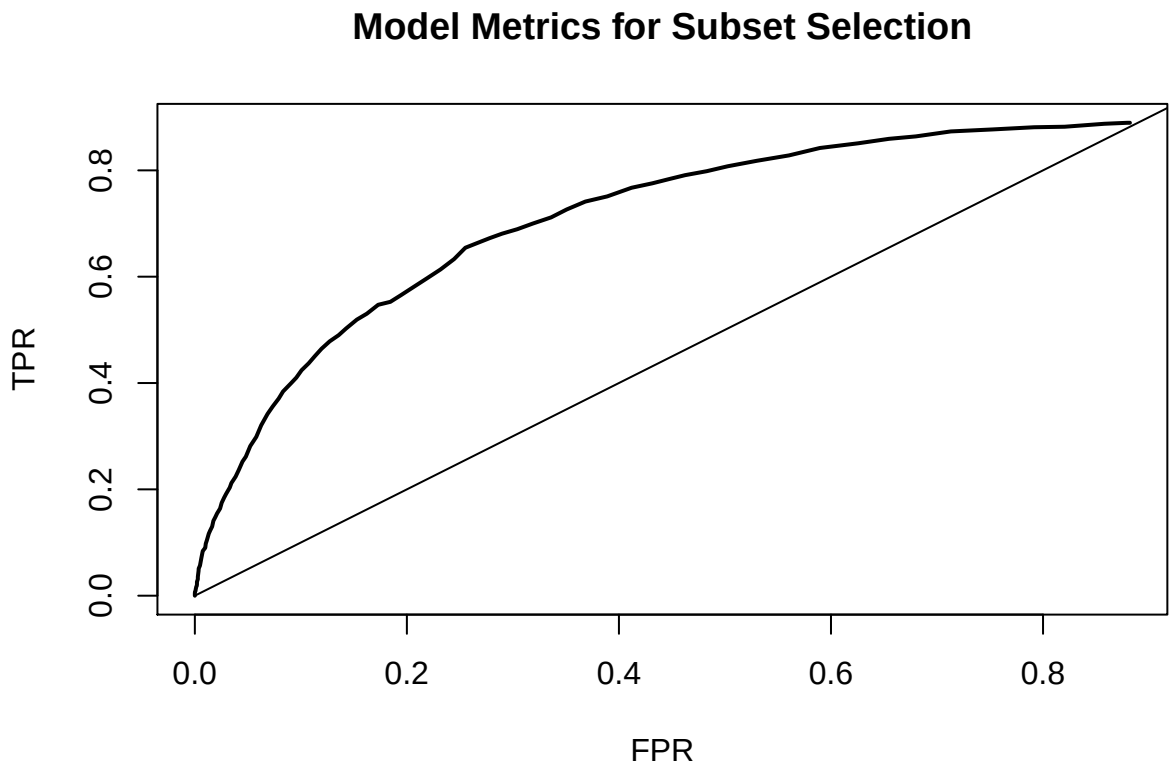
Table 19:

| | TARGET_FLAG |
|---|---|
| Constant | $-1.135^{***}$ (0.185) |
| KIDSDRIV | $0.469^{***}$ (0.052) |
| I(INCOMEˆ(1/4)) | $-0.085^{***}$ (0.010) |
| MSTATUSYes | $-0.748^{***}$ (0.057) |
| JOBClerical | 0.148 (0.099) |
| JOBDoctor | $-0.972^{***}$ (0.218) |
| JOBHome Maker | $-0.422^{***}$ (0.156) |
| JOBLawyer | $-0.467^{***}$ (0.121) |
| JOBManager | $-0.955^{***}$ (0.096) |
| JOBProfessional | $-0.398^{***}$ (0.099) |
| JOBStudent | $-0.343^{**}$ (0.144) |
| CAR_USEPrivate | $-0.725^{***}$ (0.070) |
| REVOKEDYes | $0.753^{***}$ (0.078) |
| MVR_PTS | $0.151^{***}$ (0.012) |
| URBANICITYHighly Urban/ Urban | $2.283^{***}$ (0.109) |
| *N* | 8,161 |
| Log Likelihood | −3,839.185 |
| Akaike Inf. Crit. | 7,708.369 |

*Notes:* $^{***}$Significant at the 1 percent level.
$^{**}$Significant at the 5 percent level.
$^{*}$Significant at the 10 percent level.

### 5.3.3 Model Metrics for Subset Selection

We first use an established threshold of .50 to determine our best possible threshold.




| | Act-Pos | Act-Neg |
|---|---|---|
| Pred-Pos | 736 | 1417 |
| Pred-Neg | 412 | 5596 |

```
## Error in inherits(x, "list"): object 'dfSubModel' not found
```

### 5.3.4 Multicollinearity for Subset Selection

We square `GVIF^(1/(2*Df))` in order to use the VIF threshold of 4 for multicollinearity. Here in our subset selection model we find that no variable exceeds our pre-established threshold of 4 for multicollinearity.

| rn | GVIF | Df | GVIF^(1/(2*Df)) | Adjusted_GVIF |
|---|---|---|---|---|
| KIDSDRIV | 1.022204 | 1 | 1.011041 | 1.022204 |
| I(INCOME^(1/4)) | 3.012788 | 1 | 1.735739 | 3.012788 |
| MSTATUS | 1.026972 | 1 | 1.013396 | 1.026972 |
| JOB | 4.435051 | 7 | 1.112262 | 1.237126 |
| CAR_USE | 1.491887 | 1 | 1.221428 | 1.491887 |
| REVOKED | 1.004182 | 1 | 1.002089 | 1.004182 |
| MVR_PTS | 1.011248 | 1 | 1.005608 | 1.011248 |
| URBANICITY | 1.084428 | 1 | 1.041359 | 1.084428 |

| rn | GVIF | Df | GVIFˆ(1/(2*Df)) | Adjusted_GVIF |
|---|---|---|---|---|

## 5.4 Linear Regression 1 - Backwards Selection

TARGET_AMT ~ KIDSDRIV + I(INCOMEˆ(1/4)) + PARENT1 + I(sqrt(HOME_VAL)) + MSTATUS + SEX + EDUCATION + JOB + I(sqrt(TRAVTIME)) + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY

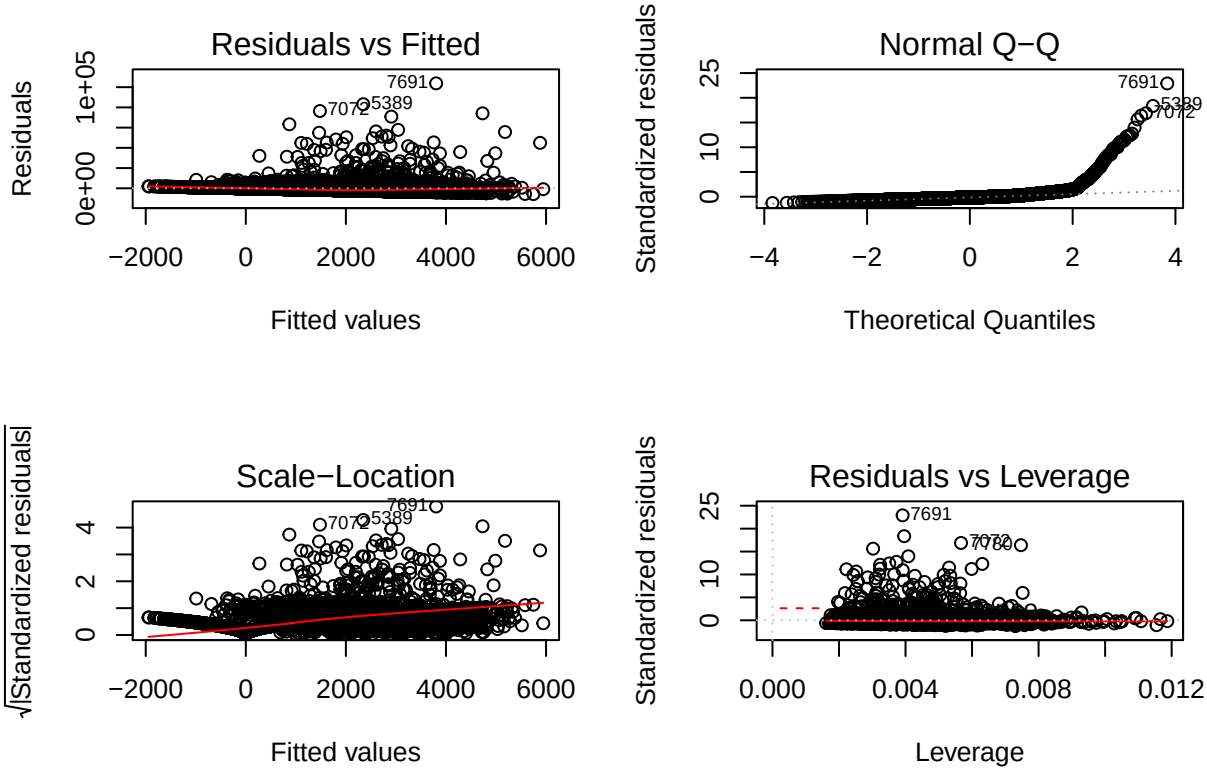
Call: lm(formula = formula(bkFitStep), data = wimputedDfTr)

Residuals: Min 1Q Median 3Q Max -5747 -1682 -765 351 103783

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.551e+02 4.837e+02 0.321 0.748439
KIDSDRIV 3.737e+02 1.020e+02 3.663 0.000250 ***I(INCOMEˆ(1/4)) -4.261e+01 1.840e+01 -2.316 0.020589***
**PARENT1Yes 6.517e+02 1.767e+02 3.688 0.000228** *I(sqrt(HOME_VAL)) -5.126e-01 3.281e-01 -1.562 0.118249*
*MSTATUSYes -4.721e+02 1.456e+02 -3.243 0.001187* ***SEXM 3.545e+02 1.610e+02 2.202 0.027689 ****
***EDUCATIONHigh School 1.798e+02 1.588e+02 1.133 0.257447***
***EDUCATIONLess Than High School 2.754e+02 2.061e+02 1.337 0.181416***
***EDUCATIONMasters 3.592e+02 2.090e+02 1.719 0.085678 .***
***EDUCATIONPhD 6.527e+02 2.683e+02 2.433 0.014998 ****
***JOBClerical 5.434e+01 1.911e+02 0.284 0.776135***
***JOBDoctor -1.228e+03 4.285e+02 -2.867 0.004160*** *JOBHome Maker -3.168e+02 2.988e+02 -1.060 0.289093*
*JOBLawyer -3.461e+02 3.047e+02 -1.136 0.256026*
*JOBManager -1.036e+03 2.248e+02 -4.610 4.08e-06* **JOBProfessional 1.103e+01 2.107e+02 0.052 0.958234**
**JOBStudent -4.657e+02 2.738e+02 -1.701 0.089048 .**
**I(sqrt(TRAVTIME)) 1.314e+02 3.497e+01 3.759 0.000172** *CAR_USEPrivate -8.244e+02 1.615e+02 -5.104 3.40e-07* **BLUEBOOK 1.350e-02 8.538e-03 1.582 0.113721**
**TIF -4.772e+01 1.217e+01 -3.922 8.86e-05** *CAR_TYPEPanel Truck 2.837e+02 2.758e+02 1.028 0.303764*
*CAR_TYPEPickup 3.976e+02 1.705e+02 2.332 0.019724*
CAR_TYPESports Car 1.036e+03 2.163e+02 4.787 1.72e-06 ***CAR_TYPESUV 7.750e+02 1.784e+02 4.343 1.42e-05*** CAR_TYPEVan 5.193e+02 2.126e+02 2.443 0.014591 *
CLM_FREQ 1.062e+02 4.879e+01 2.177 0.029497 *
REVOKEDYes 4.305e+02 1.549e+02 2.780 0.005448 ** MVR_PTS 1.702e+02 2.580e+01 6.597 4.46e-11
***CAR_AGE -2.788e+01 1.320e+01 -2.112 0.034751***
**URBANICITYHighly Urban/ Urban 1.676e+03 1.391e+02 12.046 < 2e-16** * — Signif. codes: 0 *‘’ 0.001 ’’ 0.01 ”* 0.05 ‘.’ 0.1 ” 1

Residual standard error: 4541 on 8129 degrees of freedom Multiple R-squared: 0.07161, Adjusted R-squared: 0.06807 F-statistic: 20.23 on 31 and 8129 DF, p-value: < 2.2e-16

### 5.4.1 Linear Regression - Backwards Selection Model Metrics




TARGET_AMT ~ MVR_PTS + URBANICITY + JOB + MSTATUS + CAR_USE + KIDSDRIV + CAR_TYPE + TIF + I(sqrt(TRAVTIME)) + PARENT1 + I(INCOMEˆ(1/4)) + REVOKED + CAR_AGE + CLM_FREQ + SEX + EDUCATION + BLUEBOOK + I(sqrt(HOME_VAL))

Call: lm(formula = formula(fwdFitstep), data = wimputedDfTr)

Residuals: Min 1Q Median 3Q Max -5747 -1682 -765 351 103783

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.551e+02 4.837e+02 0.321 0.748439
MVR_PTS 1.702e+02 2.580e+01 6.597 4.46e-11 ***URBANICITYHighly Urban/ Urban 1.676e+03 1.391e+02 12.046 < 2e-16*** JOBClerical 5.434e+01 1.911e+02 0.284 0.776135
JOBDoctor -1.228e+03 4.285e+02 -2.867 0.004160 ** JOBHome Maker -3.168e+02 2.988e+02 -1.060 0.289093
JOBLawyer -3.461e+02 3.047e+02 -1.136 0.256026
JOBManager -1.036e+03 2.248e+02 -4.610 4.08e-06 ***JOBProfessional 1.103e+01 2.107e+02 0.052 0.958234 JOBStudent -4.657e+02 2.738e+02 -1.701 0.089048 .***
***MSTATUSYes -4.721e+02 1.456e+02 -3.243 0.001187*** *CAR_USEPrivate -8.244e+02 1.615e+02 -5.104 3.40e-07* **KIDSDRIV 3.737e+02 1.020e+02 3.663 0.000250** *CAR_TYPEPanel Truck 2.837e+02 2.758e+02 1.028 0.303764*
*CAR_TYPEPickup 3.976e+02 1.705e+02 2.332 0.019724*
CAR_TYPESports Car 1.036e+03 2.163e+02 4.787 1.72e-06 ***CAR_TYPESUV 7.750e+02 1.784e+02 4.343 1.42e-05*** CAR_TYPEVan 5.193e+02 2.126e+02 2.443 0.014591 *
TIF -4.772e+01 1.217e+01 -3.922 8.86e-05 ***I(sqrt(TRAVTIME)) 1.314e+02 3.497e+01 3.759 0.000172*** PARENT1Yes 6.517e+02 1.767e+02 3.688 0.000228 ***I(INCOMEˆ(1/4)) -4.261e+01 1.840e+01 -2.316 0.020589***

**REVOKEDYes 4.305e+02 1.549e+02 2.780 0.005448** CAR_AGE -2.788e+01 1.320e+01 -2.112 0.034751 *
CLM_FREQ 1.062e+02 4.879e+01 2.177 0.029497 *
SEXM 3.545e+02 1.610e+02 2.202 0.027689 *
EDUCATIONHigh School 1.798e+02 1.588e+02 1.133 0.257447
EDUCATIONLess Than High School 2.754e+02 2.061e+02 1.337 0.181416
EDUCATIONMasters 3.592e+02 2.090e+02 1.719 0.085678 .
EDUCATIONPhD 6.527e+02 2.683e+02 2.433 0.014998 *
BLUEBOOK 1.350e-02 8.538e-03 1.582 0.113721
I(sqrt(HOME_VAL)) -5.126e-01 3.281e-01 -1.562 0.118249
— Signif. codes: 0 ‘’ ***0.001*** ” *0.01* ” 0.05 ‘.’ 0.1 ” 1

Residual standard error: 4541 on 8129 degrees of freedom Multiple R-squared: 0.07161, Adjusted R-squared: 0.06807 F-statistic: 20.23 on 31 and 8129 DF, p-value: < 2.2e-16