Critical Thinking Group 4 - HW4- Auto Insurance

Sreejaya, Suman, Vuthy

November 7, 2016

Overview

The purpose of this project is 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 multiple linear regression and binary logistic regression models. Below is a short description of the variables in the dataset.

-INDEX: Identification Variable (do not use) None
-TARGET_FLAG: Was Car in a crash? 1=YES 0=NO None
-TARGET_AMT: If car was in a crash, what was the cost None
-AGE: Age of Driver Very
young people tend to be risky. Maybe very old people also.
-BLUEBOOK: Value of Vehicle
Unknown effect on probability of collision, but probably effect the payout if there is a crash
-CAR_AGE: Vehicle Age
Unknown effect on probability of collision, but probably effect the payout if there is a crash
-CAR_TYPE: Type of Car
Unknown effect on probability of collision, but probably effect the payout if there is a crash
-CAR_USE: Vehicle Use
Commercial vehicles are driven more, so might increase probability of collision
-CLM_FREQ: # Claims (Past 5 Years)
The more claims you filed in the past, the more you are likely to file in the future
-EDUCATION: Max Education Level
Unknown effect, but in theory more educated people tend to drive more safely
-HOMEKIDS: # Children at Home
Unknown effect
-HOME_VAL: Home Value In theory,
home owners tend to drive more responsibly
-INCOME: Income In theory,
rich people tend to get into fewer crashes
-JOB: Job Category In theory,
white collar jobs tend to be safer
-KIDSDRIV: # Driving Children When teenagers drive your car,
you are more likely to get into crashes
-MSTATUS: Marital Status In theory,
married people drive more safely
-MVR_PTS: Motor Vehicle Record Points
If you get lots of traffic tickets, you tend to get into more crashes
-OLDCLAIM: Total Claims (Past 5 Years)
If your total payout over the past five years was high, this suggests future payouts will be high
-PARENT1: Single Parent Unknown effect
-RED_CAR: A Red Car
Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
-REVOKED: License Revoked (Past 7 Years)
If your license was revoked in the past 7 years, you probably are a more risky driver.
-SEX: Gender
Urban legend says that women have less crashes then men. Is that true?
-TIF: Time in Force

People who have been customers for a long time are usually more safe.
-TRAVTIME: Distance to Work
Long drives to work usually suggest greater risk
-URBANICITY: Home/Work Area
Unknown
-YOJ: Years on Job
People who stay at a job for a long time are usually more safe

Dataset
Crime - Training data
Crime - Evaluation Data

Data Exploration

The dataset contains 8000 observations and 26 variables. Each record has two response variables. **TARGET_FLAG** and **TARGET_AMT** Below is a glimpse of the data. A quick look indicates that chas and target might be classification variables.

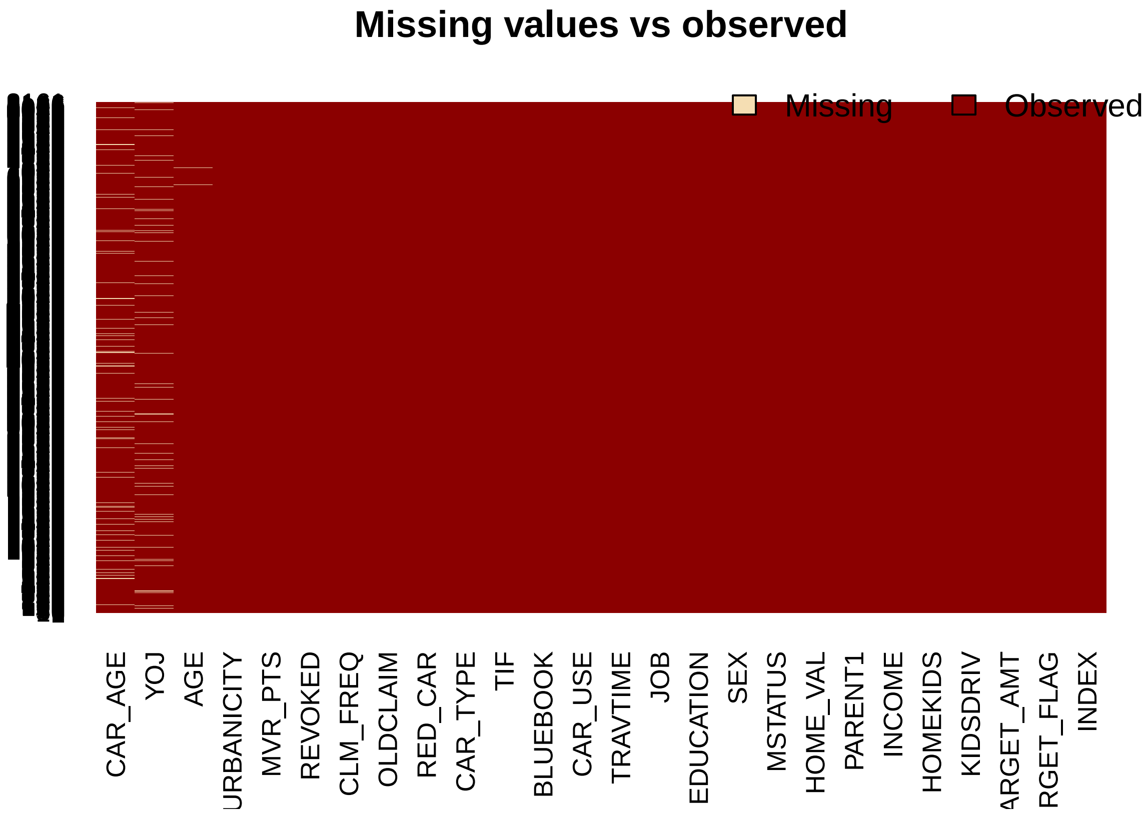
```
## Observations: 8,161
## Variables: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 1...
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0,...
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000...
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55...
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 3, 0,...
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, ...
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", "", "$114,986", "...
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "...
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,92...
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes"...
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z...
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High Sc...
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Bl...
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25,...
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Pr...
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,00...
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6...
## $ CAR_TYPE    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV",...
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes"...
## $ OLDCLAIM    <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", ...
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0,...
## $ REVOKED     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", ...
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0...
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5,...
## $ URBANICITY  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Hig...
```

```
##      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
##  Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      SEX             EDUCATION            JOB              TRAVTIME
##  Length:8161        Length:8161        Length:8161        Min.   :  5.00
##  Class :character   Class :character   Class :character   1st Qu.: 22.00
##  Mode  :character   Mode  :character   Mode  :character   Median : 33.00
##                                                           Mean   : 33.49
##                                                           3rd Qu.: 44.00
##                                                           Max.   :142.00
##
##    CAR_USE            BLUEBOOK             TIF          CAR_TYPE
##  Length:8161        Length:8161        Min.   : 1.000   Length:8161
##  Class :character   Class :character   1st Qu.: 1.000   Class :character
##  Mode  :character   Mode  :character   Median : 4.000   Mode  :character
##                                        Mean   : 5.351
##                                        3rd Qu.: 7.000
##                                        Max.   :25.000
##
##    RED_CAR            OLDCLAIM           CLM_FREQ       REVOKED
##  Length:8161        Length:8161        Min.   :0.0000   Length:8161
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.7986
##                                        3rd Qu.:2.0000
##                                        Max.   :5.0000
##
##     MVR_PTS          CAR_AGE        URBANICITY
##  Min.   : 0.000   Min.   :-3.000   Length:8161
##  1st Qu.: 0.000   1st Qu.: 1.000   Class :character
##  Median : 1.000   Median : 8.000   Mode  :character
##  Mean   : 1.696   Mean   : 8.328
##  3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :13.000   Max.   :28.000
##                   NA's   :510
```

Taking a closer look at the data with summary statistics, we can see that PARENT1, SEX, MSTATUS, CAR_USE, RED_CAR, REVOKED, URBANICITY should be converted to factors.

Visually assessing missing values:

The Amelia package has a plotting function missmap() that will plot the dataset and highlight missing values:



There are missing values in CAR__AGE and YOJ.