Data 621 - Homework 4

Group 2: William Aiken, Donald Butler, Michael Ippolito, Bharani Nittala, and Leticia Salazar

November 6, 2022

Contents

Overview:

In this homework assignment, you will explore, analyze and model a data set containing approximately 8,000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

Objective:

Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

Description:

Below is a short description of the variables of interest in the data set:

VARIABLE NAME:	DEFINITION:	THEORETICAL EFFECT:
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1 = YES 2 = NO	None

VARIABLE NAME:	DEFINITION:	THEORETICAL EFFECT:
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	Martial 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

VARIABLE NAME:	DEFINITION:	THEORETICAL EFFECT:
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

---

Load Libraries:

These are the libraries used to explore, prepare, analyze and build our models

```
library(tidyverse)
library(caret)
library(pROC)
library(corrplot)
library(GGally)
library(psych)
library(car)
library(kableExtra)
library(gridExtra)
library(performance)
library(faraway)
library(jtools)
```

Load Data set:

We have included the original data sets in our GitHub account and read from this location. Our training data set includes 8,161 records and 26 variables.

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", "", "$114,986", "$125,301~
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No~
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "$0"~
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", "Yes",~
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", "M"~
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High School", "~
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Blue Colla~
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Private", ~
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000", "$17~
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV", "Sports~
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes", "no", ~
## $ OLDCLAIM    <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0", "$~
```

```
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "N~
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly Urba~
```

Data Exploration:

For insight on the data we use the `summary()` function on the train dataset:

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

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

The following dummy variables are done to both the training and evaluation data set and only showing the results for the training data.

PARENT1 The *PARENT1* variable has two values, Yes and No, to indicate if the observation is a single parent. We will construct a dummy variable *SingleParent* = 1 if *PARENT1* = Yes.

```
##   PARENT1    n
## 1      No 7084
## 2     Yes 1077
```

SEX The *SEX* variable has two values, M and z_F. We will create a dummy variable *Male* = 1 if *SEX* = M.

```
##   SEX    n
## 1   M 3786
## 2 z_F 4375
```

MSTATUS The variable *MSTATUS* has two values, Yes and z_No, to indicate the marital status. We will create a dummy variable *Married* = 1 if MSTATUS = Yes.

```
##   MSTATUS    n
## 1     Yes 4894
## 2    z_No 3267
```

EDUCATION The *EDUCATION* variable takes on 5 values ranging from less than high school through PHD. We will construct dummy variables: *HighSchool*, *Bachelors*, *Masters*, *PHD*, to indicate the highest level of education completed.

```
##        EDUCATION    n
## 1   <High School 1203
## 2      Bachelors 2242
## 3        Masters 1658
## 4            PhD  728
## 5 z_High School 2330
```

JOB The *JOB* variable takes on 8 values. The *JOB* variable has 526 missing values, so we will construct dummy variables for all 8 values assuming the missing values are not one of the listed professions. The dummy variables we will create are: *Clerical*, *Doctor*, *HomeMaker*, *Lawyer*, *Manager*, *Professional*, *Student*, and *BlueCollar*.

```
##              JOB    n
## 1                 526
## 2       Clerical 1271
## 3         Doctor  246
## 4     Home Maker  641
## 5         Lawyer  835
## 6        Manager  988
## 7   Professional 1117
## 8        Student  712
## 9 z_Blue Collar 1825
```

CAR_USE The *CAR_USE* variable has two values, Commercial and Private. We will construct a dummy variable *Commercial* = 1 if Commercial.

```
##      CAR_USE    n
## 1 Commercial 3029
## 2    Private 5132
```

CAR_TYPE The *CAR_TYPE* variable takes on 6 values. We will create dummy variables; *Minivan*, *PanelTruck*, *Pickup*, *SportsCar*, and *Van*.

```
##      CAR_TYPE    n
## 1     Minivan 2145
## 2 Panel Truck  676
## 3      Pickup 1389
## 4  Sports Car  907
## 5         Van  750
## 6       z_SUV 2294
```

RED_CAR The *RED_CAR* variable has two values, yes and no. We will create a dummy variable *RedCar* = 1 if *RED_CAR* = yes.

```
##   RED_CAR    n
## 1      no 5783
## 2     yes 2378
```

REVOKED The *REVOKED* variable has two values, Yes and No. We will create a dummy variable *DLRevoked* = 1 if *REVOKED* = Yes.

```
##   REVOKED    n
## 1      No 7161
## 2     Yes 1000
```

URBANICITY The *URBANICITY* variable has two values, Highly Urban/ Urban and z_Highly Rural/ Rural. We will create a dummy variable *Urban* = 1 if *URBANICITY* = Highly Urban/ Urban.

```
##              URBANICITY    n
## 1   Highly Urban/ Urban 6492
## 2 z_Highly Rural/ Rural 1669
```

Data Preparation:

Performed to both the training and evaluation data sets.

Data Cleaning Function

- The attributes `BLUEBOOK`, `HOME_VAL`, `INCOME`, and`OLDCLAIM` are dollar amounts stored as characters. Need to convert to int.
- Variables with NA: `AGE` (6), `YOJ` (454), `CAR_AGE` (510)
- Consider creating `AGE` groups Under25 and Over65 to account for young and older drivers.
- Consider creating `CAR_AGE` groups to identify new cars. One observation has a `CAR_AGE` = -3, which shouldn't be possible.
- Consider creating `YOJ` (Year on Job) groups to identify job stability; Over5years etc.

```
##   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     NA   306251
## 5     6           0          0        0  50        0  NA 114986   243925
## 6     7           1       2946        0  34        1  12 125301        0
```

```
##   TRAVTIME BLUEBOOK TIF OLDCLAIM CLM_FREQ MVR_PTS CAR_AGE SingleParent Male
## 1       14    14230  11     4461        2       3      18            0    1
## 2       22    14940   1        0        0       0       1            0    1
## 3        5     4010   4    38690        2       3      10            0    0
## 4       32    15440   7        0        0       0       6            0    1
## 5       36    18000   1    19217        2       3      17            0    0
## 6       46    17430   1        0        0       0       7            1    0
##   Married HighSchool Bachelors Masters PHD Clerical Doctor HomeMaker Lawyer
## 1       0          0         0       0   1        0      0         0      0
## 2       0          1         0       0   0        0      0         0      0
## 3       1          1         0       0   0        1      0         0      0
## 4       1          0         0       0   0        0      0         0      0
## 5       1          0         0       0   1        0      1         0      0
## 6       0          0         1       0   0        0      0         0      0
##   Manager Professional Student BlueCollar Commercial Minivan PanelTruck Pickup
## 1       0            1       0          0          0       1          0      0
## 2       0            0       0          1          1       1          0      0
## 3       0            0       0          0          0       0          0      0
## 4       0            0       0          1          0       1          0      0
## 5       0            0       0          0          0       0          0      0
## 6       0            0       0          1          1       0          0      0
##   SportsCar Van RedCar DLRevoked Urban
## 1         0   0      1         0     1
## 2         0   0      1         0     1
## 3         0   0      0         0     1
## 4         0   0      1         0     1
## 5         0   0      0         1     1
## 6         1   0      0         0     1
```

Model Building:

We will be building five different models; two multiple linear regression models and three binary logistic regression models.

Model 1

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = cleandf, na.action = na.omit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6493   -530    -58    273  79064
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.338e+02  5.274e+02  -0.254   0.7998
## INDEX        -3.572e-03  1.648e-02  -0.217   0.8285
## TARGET_FLAG   5.643e+03  1.266e+02  44.578  < 2e-16 ***
## KIDSDRIV     -1.995e+01  1.107e+02  -0.180   0.8570
## AGE           1.362e+00  6.902e+00   0.197   0.8436
```

```
## HOMEKIDS      4.562e+01  6.364e+01   0.717   0.4734
## YOJ           6.837e+00  1.455e+01   0.470   0.6383
## INCOME       -3.196e-03  1.851e-03  -1.727   0.0842 .
## HOME_VAL      3.297e-04  5.908e-04   0.558   0.5768
## TRAVTIME      1.000e+00  3.155e+00   0.317   0.7512
## BLUEBOOK      3.356e-02  8.341e-03   4.023 5.81e-05 ***
## TIF          -2.065e-01  1.189e+01  -0.017   0.9861
## OLDCLAIM      9.694e-03  7.215e-03   1.344   0.1791
## CLM_FREQ     -7.767e+01  5.350e+01  -1.452   0.1466
## MVR_PTS       4.384e+01  2.534e+01   1.730   0.0836 .
## CAR_AGE      -2.359e+01  1.240e+01  -1.903   0.0571 .
## SingleParent  7.409e+01  1.958e+02   0.378   0.7051
## Male          3.504e+02  1.768e+02   1.981   0.0476 *
## Married      -2.734e+02  1.446e+02  -1.891   0.0587 .
## HighSchool   -2.050e+02  1.671e+02  -1.227   0.2197
## Bachelors     2.939e+01  2.010e+02   0.146   0.8837
## Masters       8.780e+01  2.957e+02   0.297   0.7666
## PHD           5.291e+02  3.542e+02   1.494   0.1352
## Clerical     -2.792e+02  3.340e+02  -0.836   0.4032
## Doctor       -5.451e+02  3.982e+02  -1.369   0.1711
## HomeMaker    -1.876e+02  3.606e+02  -0.520   0.6029
## Lawyer        1.481e+01  2.893e+02   0.051   0.9592
## Manager      -2.123e+02  2.842e+02  -0.747   0.4550
## Professional  1.881e+02  3.016e+02   0.624   0.5329
## Student      -3.422e+02  3.700e+02  -0.925   0.3550
## BlueCollar    7.685e+01  3.148e+02   0.244   0.8072
## Commercial    1.424e+01  1.615e+02   0.088   0.9297
## Minivan      -2.894e+02  1.730e+02  -1.673   0.0944 .
## PanelTruck   -1.881e+02  3.298e+02  -0.570   0.5685
## Pickup       -2.443e+02  1.932e+02  -1.265   0.2060
## SportsCar     1.011e+02  1.743e+02   0.580   0.5617
## Van          -2.581e+02  2.587e+02  -0.998   0.3185
## RedCar       -4.581e+01  1.450e+02  -0.316   0.7521
## DLRevoked    -3.023e+02  1.707e+02  -1.771   0.0766 .
## Urban        -2.730e+01  1.412e+02  -0.193   0.8466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3931 on 6408 degrees of freedom
##   (1713 observations deleted due to missingness)
## Multiple R-squared:  0.2945, Adjusted R-squared:  0.2902
## F-statistic: 68.58 on 39 and 6408 DF,  p-value: < 2.2e-16
```

Model 2

Model 3

Model 4

Model 5

Select Models: