DATA 621: BUSINESS ANALYTICS AND DATA MINING HOMEWORK#4: LOGISTIC REGRESSION

Group 2 - Gabriel Campos, Melissa Bowman, Alexander Khaykin, & Jennifer Abinette

Last edited November 22, 2023

Contents

1 Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 8000 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.

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:

Variable Names	Definition	Theoretical Effect
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	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	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

1.1 Deliverables

- A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.
- Assigned predictions (probabilities, classifications, cost) for the evaluation data set. Use 0.5 threshold.
- Include your R statistical programming code in an Appendix.

1.2 Write Up:

1.2.1 1. DATA EXPLORATION (25 Points)

Describe the size and the variables in the insurance training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

1.2.2 2. DATA PREPARATION (25 Points)

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

a. Fix missing values (maybe with a Mean or Median value)
b. Create flags to suggest if a variable was missing
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root (or use Box-Cox)
e. Combine variables (such as ratios or adding or multiplying) to create new variables

1.2.3 3. BUILD MODELS (25 Points)

Using the training data set, build at least two different multiple linear regression models and three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? For example, if a person has a lot of traffic tickets, you would reasonably expect that person to have more car crashes. If the coefficient is negative (suggesting that the person is a safer driver), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

1.2.4 4. SELECT MODELS (25 Points)

Decide on the criteria for selecting the best multiple linear regression model and the best binary logistic regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models.

For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. 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 you 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.

2 Import Data

```
df_insur_eval <-
  read.csv(paste0(url_git,"insurance-evaluation-data.csv"))

head(df_insur_eval,n=10)
```

```
##    INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1
## 1      3          NA         NA        0  48        0  11  $52,881      No
## 2      9          NA         NA        1  40        1  11  $50,815     Yes
## 3     10          NA         NA        0  44        2  12  $43,486     Yes
## 4     18          NA         NA        0  35        2  NA  $21,204     Yes
## 5     21          NA         NA        0  59        0  12  $87,460      No
## 6     30          NA         NA        0  46        0  14               No
## 7     31          NA         NA        0  60        0  12  $37,940      No
## 8     37          NA         NA        0  54        0  12  $33,212      No
```

```
## 9     39          NA          NA         2  36         2  12 $130,540     Yes
## 10    47          NA          NA         0  50         0   8 $167,469      No
##    HOME_VAL MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE
## 1        $0    z_No   M     Bachelors       Manager       26    Private
## 2        $0    z_No   M z_High School       Manager       21    Private
## 3        $0    z_No z_F z_High School z_Blue Collar       30 Commercial
## 4        $0    z_No   M z_High School      Clerical       74    Private
## 5        $0    z_No   M z_High School       Manager       45    Private
## 6  $207,519     Yes   M     Bachelors  Professional        7 Commercial
## 7  $182,739     Yes z_F z_High School z_Blue Collar       16 Commercial
## 8  $158,432     Yes   M  <High School z_Blue Collar       27 Commercial
## 9  $344,195    z_No z_F     Bachelors z_Blue Collar        5 Commercial
## 10       $0    z_No z_F           PhD        Doctor       22    Private
##    BLUEBOOK TIF    CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1   $21,970   1         Van     yes       $0        0      No       2      10
## 2   $18,930   6     Minivan      no   $3,295        1      No       2       1
## 3    $5,900  10       z_SUV      no       $0        0      No       0      10
## 4    $9,230   6      Pickup      no       $0        0     Yes       0       4
## 5   $15,420   1     Minivan     yes  $44,857        2      No       4       1
## 6   $25,660   1 Panel Truck      no   $2,119        1      No       2      12
## 7   $11,290   1  Sports Car      no       $0        0      No       0       1
## 8   $24,000   4 Panel Truck      no       $0        0      No       5      NA
## 9   $27,200   4     Minivan      no       $0        0      No       0       9
## 10  $34,150   4  Sports Car      no       $0        0      No       3       1
##              URBANICITY
## 1    Highly Urban/ Urban
## 2    Highly Urban/ Urban
## 3  z_Highly Rural/ Rural
## 4  z_Highly Rural/ Rural
## 5    Highly Urban/ Urban
## 6    Highly Urban/ Urban
## 7    Highly Urban/ Urban
## 8    Highly Urban/ Urban
## 9  z_Highly Rural/ Rural
## 10   Highly Urban/ Urban
```

```r
df_insur_train <-
  read.csv(paste0(url_git,"insurance_training_data.csv"))

head(df_insur_train,n=10)
```

```
##    INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1
## 1      1           0          0        0  60        0  11  $67,349      No
## 2      2           0          0        0  43        0  11  $91,449      No
## 3      4           0          0        0  35        1  10  $16,039      No
## 4      5           0          0        0  51        0  14               No
## 5      6           0          0        0  50        0  NA $114,986      No
## 6      7           1       2946        0  34        1  12 $125,301     Yes
## 7      8           0          0        0  54        0  NA  $18,755      No
## 8     11           1       4021        1  37        2  NA $107,961      No
## 9     12           1       2501        0  34        0  10  $62,978      No
## 10    13           0          0        0  50        0   7 $106,952      No
##    HOME_VAL MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE
## 1        $0    z_No   M           PhD  Professional       14    Private
```

```
## 2  $257,252    z_No   M z_High School z_Blue Collar       22 Commercial
## 3  $124,191     Yes z_F z_High School      Clerical        5    Private
## 4  $306,251     Yes   M  <High School z_Blue Collar       32    Private
## 5  $243,925     Yes z_F           PhD        Doctor       36    Private
## 6        $0    z_No z_F     Bachelors z_Blue Collar       46 Commercial
## 7               Yes z_F  <High School z_Blue Collar       33    Private
## 8  $333,680     Yes   M     Bachelors z_Blue Collar       44 Commercial
## 9        $0    z_No z_F     Bachelors      Clerical       34    Private
## 10       $0    z_No   M     Bachelors  Professional       48 Commercial
##    BLUEBOOK TIF   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1   $14,230  11    Minivan     yes   $4,461        2      No       3      18
## 2   $14,940   1    Minivan     yes       $0        0      No       0       1
## 3    $4,010   4      z_SUV      no  $38,690        2      No       3      10
## 4   $15,440   7    Minivan     yes       $0        0      No       0       6
## 5   $18,000   1      z_SUV      no  $19,217        2     Yes       3      17
## 6   $17,430   1 Sports Car      no       $0        0      No       0       7
## 7    $8,780   1      z_SUV      no       $0        0      No       0       1
## 8   $16,970   1        Van     yes   $2,374        1     Yes      10       7
## 9   $11,200   1      z_SUV      no       $0        0      No       0       1
## 10  $18,510   7        Van      no       $0        0      No       1      17
##               URBANICITY
## 1    Highly Urban/ Urban
## 2    Highly Urban/ Urban
## 3    Highly Urban/ Urban
## 4    Highly Urban/ Urban
## 5    Highly Urban/ Urban
## 6    Highly Urban/ Urban
## 7    Highly Urban/ Urban
## 8    Highly Urban/ Urban
## 9    Highly Urban/ Urban
## 10 z_Highly Rural/ Rural
```